ITT

QUANTERION
SOLUTIONS INCORPORATED

School of Information Studies

Defense Technical Information Center

Managed Services for DoD-Generated Datasets

Final Report

Contract No. SPO700-98-D-4000

03 March 2010

Prepared by:

Victor Choo
ITT Corporation
Advanced Engineering & Sciences
474 Phoenix Drive
Rome, NY 13441-4911

Thomas McGibbon
Quanterion Solutions Inc.
811 Court Street
Utica, NY 13502

Carlos Villalba
Syracuse University
113 Brown Hall
Syracuse, NY 13244

Prepared for:
Defense Technical Information Center
8725 John J. Kingman Highway
Fort Belvoir, VA 22060-6218

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-03-2010	Final	06/10/2009 – 01/31/2010

4. TITLE AND SUBTITLE

Defense Technical Information Center Managed Services for DoD-Generated Datasets

5a. CONTRACT NUMBER
SPO-0700-98-D-4000

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Victor Choo, ITT Corporation
Thomas McGibbon, Quanterion Solutions Incorporated
Carlos Villalba, Syracuse University

5d. PROJECT NUMBER
DACS

5e. TASK NUMBER
DACS DO 61

5f. WORK UNIT NUMBER
NA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

ITT Corporation
474 Phoenix Drive
Rome, NY 13441

8. PERFORMING ORGANIZATION REPORT NUMBER
ITT/DACS/DO61/TR2010/01

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Technical Information Center
8725 John J. Kingman Highway
Fort Belvoir, VA 22060

10. SPONSOR/MONITOR'S ACRONYM(S)
DTIC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution statement A: Approved for Public release

13. SUPPLEMENTARY NOTES
NA

14. ABSTRACT
The objective of this effort is to examine the DoD Research, Development, Test & Evaluation (RDT&E) community's interest in establishing a capability that allows DoD users to discover and acquire datasets that are created as part of DoD RDT&E programs. In this context, the term dataset means a collection of related information and may include items such as sensor data (e.g., radar data), video data, still images, or monitored information (e.g., computer network traffic) to name a few. The data may be stored in a series of files or a database of records. DTIC has been chartered with the dissemination of technical reports generated by the DoD RDT&E community for a number of years. The inclusion of datasets, some of which are collected to generate the technical reports, is a natural enhancement of DTIC's services.

15. SUBJECT TERMS
Managed Services, Datasets, Large Datasets, Metadata, DoD Generated Datasets

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gopalakrishnan Nair
Unclassified	Unclassified	Unclassified		52	19b. TELEPHONE NUMBER *(include area code)* 703-767-9055

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

List of Figures

List of Tables

Executive Summary

The objective of the Managed Services for DoD-Generated Datasets (MSDD) effort, also referred to as the DTIC Large Dataset effort, is to establish a capability that allows Department of Defense (DoD) users to discover and acquire datasets that are created as part of DoD Research, Development, Test & Evaluation (RDT&E) programs. In this context, the term dataset means a collection of related information and may include items such as sensor data (e.g., radar data), video data, still images, or monitored information (e.g., computer network traffic) to name a few. The data may be stored in a series of files or a database of records.

The rationale for this effort is that the DoD RDT&E organizations collectively spend millions of dollars each year collecting data. The vast majority of this data is only used to support the program that collected the data. The organization that collected the data typically stores it either on their own computer systems or on some type of offline media. The data is maintained until the program ends or until disk space is needed on the hosting computer system at which point it is removed. Meanwhile, others in the research and development (R&D) community seek sample data to test hypothesis, algorithms or approaches.

The MSDD effort was initiated by Defense Technical Information Service Research group (DTIC) to investigate the DoD RDT&E community's interest in DTIC providing the MSDD service capability. An initial study was established through the Data and Analysis Center for Software (DACS) to perform this investigation. A team of personnel from ITT Corporation, Quanterion Solutions Incorporated, and Syracuse University School of Information Studies performed the effort.

Meetings were held with representatives from the Office of the Secretary of Defense (OSD), the Air Force Research Laboratory (AFRL), the Army Research Laboratory (ARL), the Naval Research Laboratory (NRL), and the High Performance Computing Modernization Program. The results indicated that all the groups recognize the need for a program such as the MSDD. They also indicated that DTIC is the right organization to spearhead the effort while leveraging the work that other groups either already have in place or are developing.

A metadata study was also performed and identified current metadata standards and datasets that exist in the community that can be leveraged by the MSDD.

The recommendations are to continue developing a program plan for the MSDD that DTIC can use to move the program forward. It is also recommended that several prototypes be developed to demonstrate various concepts. The primary prototype is to demonstrate that DTIC's Common Validation system can be leveraged by other Government agencies to assist in the administration of existing sites that have datasets available.

1.0 Introduction

The Managed Services for DoD-Generated Datasets (MSDD) initiative is a proposed effort to disseminate datasets within the DoD community to support research and development projects. This effort was initiated by Defense Technical Information Service Research group (DTIC) to investigate the Department of Defense (DoD) Research, Development, Test & Evaluation (RDT&E) community's interest in establishing managed services infrastructure to support the sharing of RDT&E generated datasets. An effort was established through the Data Analysis Center for Software (DACS) to perform this investigation. A team of personnel from ITT Corporation, Quanterion Solutions Incorporated, and Syracuse University School of Information Studies performed the effort.

1.1 Purpose

The purpose of this document is to provide a summary of the results of the initial feasibility study performed in support of the effort. The results of this study will be used as inputs to drive the next phase of the effort.

1.2 Scope

This document is limited to describing the background of the project, the results of the initial studies, and the recommendations for proceeding with future phases of the project.

1.3 Background

The objective of the MSDD effort, also referred to as the Defense Technical Information Center (DTIC) Large Dataset effort, is to establish a capability that allows DoD users to discover and acquire datasets that are created as part of DoD RDT&E programs. In this context, the term dataset means a collection of related information and may include items such as sensor data (e.g., radar data), video data, still images, or monitored information (e.g., computer network traffic) to name a few. The data may be stored in a series of files or a database of records. DTIC has been chartered with the dissemination of technical reports generated by the DoD RDT&E community for a number of years. The inclusion of datasets, some of which are collected to generate the technical reports, is a natural enhancement of DTIC's services.

The rationale for this effort is that DoD RDT&E organizations collectively spend millions of dollars each year collecting data. The vast majority of this data is only used to support the program which collected the data. The organization that collected the data typically stores it either on their own computer systems or on some type of offline media. The data is maintained until the program ends or until disk space is needed on the hosting computer system at which point it is removed. Meanwhile, others in the R&D community seek sample data to test

hypothesis, algorithms or approaches. Generally, information sharing takes place within relatively small, close knit communities-of-interest (COI). Members of those COI's exchange data between themselves, but rarely with the outsiders.

The DoD does not operate a network service to manage data sets that have potential for re-use among the scientific and technical community. The Defense Technical Information Center (DTIC) is interested in leading the establishment of services to manage DoD-generated data sets. This proposed Managed Services for DoD-Generated Datasets capability represents an extension of DTIC's current role for disseminating technical reports and documents and falls within their DoDI 3200.14 mission statement: "The DTIC shall act as a central coordinating point for DoD STI databases and systems, and investigate and demonstrate new supporting technology for those applications."

2.0 Methods and Assumptions

The project team's approach to this effort consisted of two primary task areas: (1) engaging stakeholders through face-to-face meetings and teleconferences, and (2) investigating metadata standards and potential ontologies to could support latter stages of the effort. The approaches for performing theses tasks are outlined below. The results of these tasks are described in Section 3.

2.1 Stakeholder Meetings

Meetings with stakeholders (e.g., organizations that would be either sources of data, consumers of data, or both) were held in order to determine the community's support for a program like the MSDD effort. A secondary goal was to capture as many initial requirements and/or suggestions/recommendations as possible. The project team focused on identifying relevant organizations to engage. The following service laboratories were selected as a logical starting point:

- Air Force Research Laboratory (AFRL)
- Army Research Laboratory (ARL)
- Naval Research Laboratory (NRL)

In addition, the project team capitalized on opportunities for engaging additional organizations including the Office of the Secretary of Defense (OSD) and the High Performance Computing Modernization Program (HPCMP).

Prior to engaging the stakeholders, a "MSDD Stakeholders Meeting Presentation" was developed that describes the overall objectives and rationale for the effort as well as an initial set of questions which were intended to foster open discussions. A copy of the MSDD Stakeholders Meeting Presentation is provided in Appendix I.

The individual meetings were scheduled to last between 1.5 and 3 hours. The meetings were attended by DTIC personnel, the project team, and representatives from the stakeholder organization. The meetings started with the MSDD Stakeholders Meeting Presentation which describes program objectives, potential approaches and questions for discussion. The presentation was followed by a period of open discussion. The meetings were kept as informal round table discussions. The notes of the meeting were prepared and submitted to DTIC. The information captured during the stakeholder meetings forms a significant portion of the results of this effort.

2.2 Metadata Study

A metadata study was performed in conjunction with the stakeholder meetings. The original intent of this study was to identify potential patterns in metadata that could support the use of

automated data discovery tools. The tools are intended to search the web and identify potential datasets of interest.

The initial steps of the effort identified potential datasets of interest and attempting to analyze the existing metadata. The data that was examined varied from initially investigating the dataset links on the DTIC website to several other DoD generated datasets.

3.0 Results

The effort started with creating a vision for the MSDD and sharing this vision with stakeholders to spur discussions with other the stakeholder community. This vision is captured in the MSDD Stakeholders Meeting Presentation. This presentation describes the overall rationale for the program, presents the stages required to complete the program, describes related dataset archives, and presents a series of questions for the user community to assist in facilitating the discussions. The underlying rationale for this effort is that DoD RDT&E community has a need for datasets to validate research and use in testing new algorithms, models, and hypotheses. Meanwhile, there are a number of DoD sponsored RDT&E projects that generate datasets. Typically, these datasets are only used by the project that generated them or within small communities of interest where the members exchange information based upon past working relationships. The MSDD program seeks to facilitate this transfer of information on datasets, in much the same why that technical documents are disseminated.

An initial program outline was established and provided as part of the original proposal. This plan consists of 9 stages shown in Figure 1.

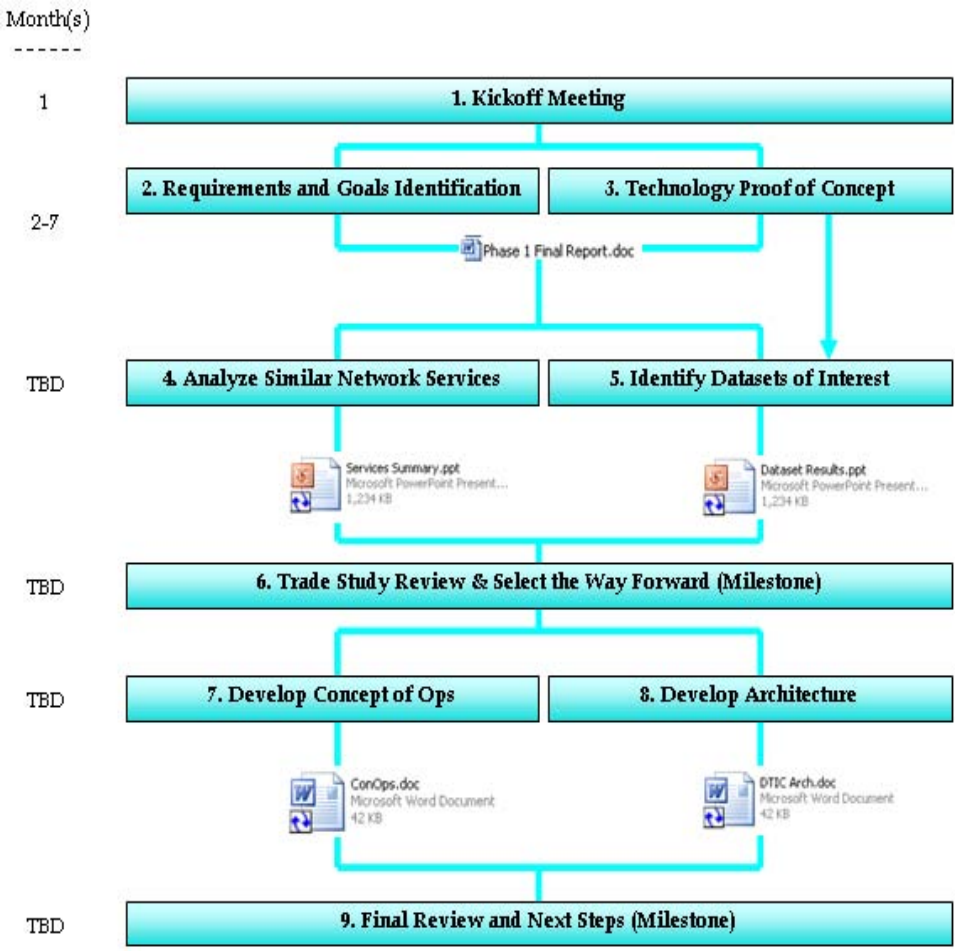


Figure 1: Initial MSDD Program Outline

The work performed under this effort supports Task 2 and Task 3. Initial difficulties in meeting with stakeholders resulted in identification of goals and suggestions, but did not yield a definitive list of requirements. The Technology Proof-of-Concept (Task 3) originally focused on investigating methods for discovering data. This task was redirected to perform an initial assessment of data ontologies and technologies that could be leveraged in identifying datasets. It was determined that it was premature to pursue this task in any depth. This is partially due to wide scope of the dataset universe, network restrictions within the DoD, and the need to narrowly focus the initial work. The latter is necessary in order to prevent becoming overwhelmed at the outset due to breadth of data and to be able to demonstrate enough depth in a particular area to be meaningful.

From an initial review of a number of similar network type services offered by other Government organizations[1], the project team recognized that DTIC could consider implementing a very broad range of services, such as indicated in Figure 2.

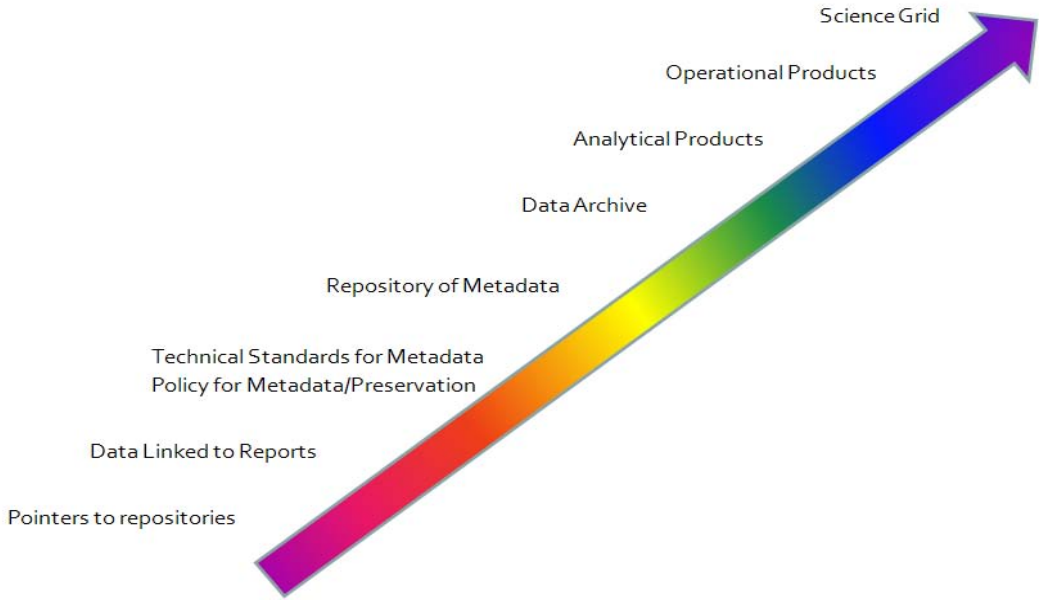


Figure 2: Potential Range of Services

For example at the lowest level, DTIC might provide a limited offering of only pointers or links to datasets. The next possible level could be expanded to offering linkages between technical reports in the DTIC collection and the datasets. As the spectrum of possible offerings increases, services offered could include a repository of metadata that describes the available datasets. The next level is to function as a data archive for orphaned datasets (e.g., those that are not longer hosted by third party sites or have exceeded the lifetime of their project), or datasets that community members are willing to share, but do not want to incur the expense and effort to host the data themselves. The archives can be expanded to include analytical tools to assist in processing the data. Finally, the data archive services could be expanded to operate with larger entities such as the Science Grid. The project team initially perceived that the desired level of service fell somewhere in the middle of the spectrum of Figure 2.

A vision of the MSDD functions was created to spur the discussion. The conceptual architecture is illustrated in Figure 3.

[1] For example: Goddard Institute for Space Studies Data and Images at http://www.giss.nasa.gov/, and the DOE's "DOE Data Explorer" at http://www.osti.gov/dataexplorer/

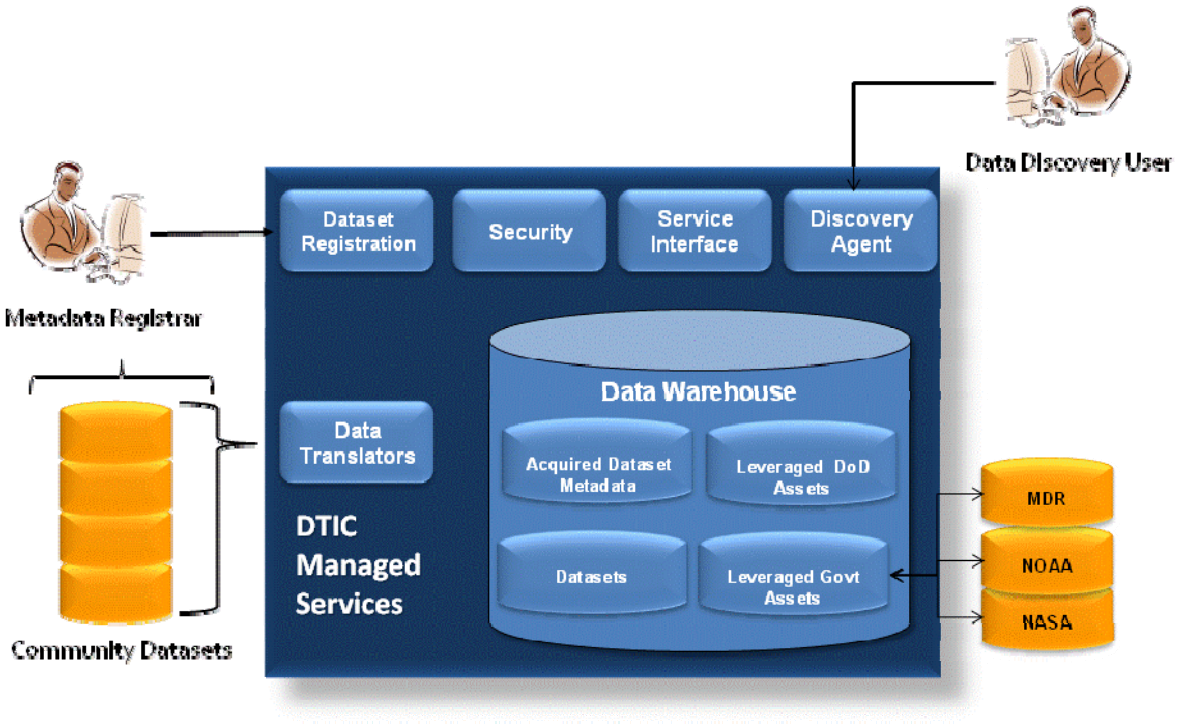


Figure 3: MSDD Conceptual Architecture

The components are described below:

- Dataset Registration – The capability allows users to register datasets and the metadata that describes the datasets. The registration database becomes a searchable inventory for the system.
- Security – The security module provides the authentication and authorization functions that control the user access for accessing and retrieving datasets.
- Service Interface – This provides an interface to external systems for exchanging security information, metadata, and data.
- Discovery Agent – The discovery agent performs several functions. It discovers data in other metadata repositories as well as actively searching the DoD web to identify new datasets.
- Community Datasets – These represent third party dataset hosts that either currently exist or are under development.
- Data Translators – These functions translate data as necessary to conform to the data standards defined by DTIC.

- Data Warehouse (Acquired Dataset Metadata) – Stores metadata that is acquired through the data registration process or acquired from third party sites.
- Data Warehouse (Datasets) – Stores orphaned datasets or those that the data owner does not want to host.
- Data Warehouse (DoD Assets) – Linkage to other DoD metadata assets or dataset lists.
- Data Warehouse (Government Assets) – Linkage to non-DoD Government metadata assets or dataset lists.

This concept is subject to change as the program becomes more defined.

3.1 Stakeholder Meeting Results

The stakeholder organizations that participated in the meetings held during this effort are listed in Appendix II.

The inputs received from the stakeholders covered a wide range of topics. The comments ranged from broad desirable features to very specific capabilities. Others were general information exchanges of ongoing projects and suggestions for proceeding. For the purposes of this report, the stakeholder's comments are divided into the following categories:

1. Capability Need/Community Support – Comments supporting the need for the program.
2. Project Synergies – Comments regarding other projects that should be investigated for synergies.
3. Dataset Needs - Capabilities and features that the MSDD should demonstrate
4. Challenges – Potential problem areas that may need to be addressed as part of the effort.
5. Datasets – A listing of potential candidate datasets identified by the stakeholders.

The following sections discuss the findings in each of the areas as well as the supporting comments.

3.1.1 Capability Need/Community Support

The majority of the stakeholders interviewed recognize the need and value of providing this type of service to the DoD RDT&E community. They recognize that as budgets shrink, the ability to reuse data can result in time/resource savings; however, the data must be in good shape and contain sufficient description such that users can quickly determine whether it addresses their needs. Currently, the sharing of datasets is performed in two ways. Some datasets are available on various DoD websites. The cost of hosting the data is typically paid for the sponsoring program. In the second method, the data sharing typically performed within specialized communities of interest. Generally, the community members know each other and are willing to

share data that they collected with each other. “Outsiders” tend to have limited views into these communities.

The stakeholders also agree that it is within DTIC’s mission to provide this type of service. In fact, it is probably a better fit under DTIC’s charter than any other organization.

The stakeholders also indicated that it is imperative that the MSDD be able to show a positive return on investment in which the amount of funding saved is substantially greater than the amount required to develop and sustain the capability. The cost of collecting the data can vary widely depending on the complexity of the data collection, amount of data, and specialty equipment required. For example, a radar experiment that requires the use of a radar system, personnel to perform the experiment, etc. can cost several million dollars to support a few days of actual data collection. Other experiments can be performed by a single person and may cost a few thousand dollars to collect.

The specific stakeholder comments regarding the need for this capability are listed in Table 1.

Table 1: Stakeholder Comments - Capability Need /Community Support

Ref#	Stakeholder Comment	Stakeholder
CS-1	DTIC should be the central DoD distribution and discovery point for this dataset capability because they have of all the authentication and information distribution capabilities available at DTIC today.	AFRL
CS-2	This is an important initiative, because the availability of datasets is becoming more limited all the time due to increases in information assurance postures. DTIC would provide an important service by performing this effort.	AFRL
CS-3	The scope of the effort is within DTIC mission statement.	AFRL, ARL
CS-4	There is a value in having the data for validating the results of a study from a scientific perspective. The NRL technical library has been asked on several occasions if they can supply the data that supports a particular report.	NRL
CS-5	DTIC should lead this type of effort for the DoD.	NRL
CS-6	Government Program Managers have been asked to provide datasets to support development efforts. This is particularly common request from university researchers.	OSD
CS-7	It is important to demonstrate the return on investment of the program, in terms of increasing the efficiency of the RDT&E development process. This includes savings of time (schedule) and labor.	OSD

3.1.2 Project Synergies

The stakeholder discussion revealed a few projects that may represent opportunities for leveraging. In particular, the OSD sponsored Joint Data Management (JDM) effort appears to have several thrusts that could benefit the MSDD. In fact, the MSDD could serve as the recipient of some of the JDM developed capabilities particularly in the metadata development areas and

the development of utility applications for working with large datasets. While the JDM has a different intended target audience, it is developing capabilities that can be leveraged by the MSDD. These leveraging opportunities need to be explored.

The Interagency Working Group on Digital Data is participating in the Networking and Information Technology Research and Development (NITRD) program (http://www.nitrd.gov/about/harnessing_power.aspx) is working on a strategy to "*promote presentation and access to digital scientific data*". DTIC participates in many of these working groups. The goal is to have the MSDD become a component that supports the NITRD strategy.

The specific stakeholder comments regarding the project synergies with the MSDD are listed in Table 2.

Table 2: Stakeholder Comments - Project Synergies

Ref#	Stakeholder Comment	Stakeholder
PS-1	The JDM has a common interest in using a centralized access control method to alleviate the need for data suppliers to handle the administrative costs associated with validating users. Leverage DTIC's existing access control system.	AFRL
PS-2	A capability such as the AFRL-led Aristotle social-networking technology could be used to allow communities of practice to comment on the goodness and validity of datasets. AFRL would like to transition to DTIC.	AFRL
PS-3	The JDM is investigating methods to improve the efficiency of processing and transporting large amounts of sensor data.	OSD
PS-4	The Networking and Information Technology Research and Development (NITRD) organization is investigating information preservation and community interaction.	OSD
PS-5	The JDM is investigating technical standards for metadata. They are currently researching scientific methods for data processing to determine what must be contained in the metadata.	OSD
PS-6	The JDM is investigating analytical products for manipulating datasets.	OSD

3.1.3 Supporting Activities

The stakeholders identified several peripheral activities that DTIC should pursue in order to support the MSDD. In particular, the stakeholders stated that DTIC should define standards for data and/or metadata. The recommended approach is to work with COI's for the domain specific detail while DTIC develops the overarching metadata. This allows users that may be starting data collection efforts to identify the information that needs to be collected up front such that it can be accounted for in terms of effort. In addition, it allows the researchers to capture the information before it becomes lost or forgotten. The guidebooks are intended to assist users in data collection methods and documentation procedures. The guidebooks could also serve as an electronic notebook for use in storing all project related information.

One group also suggested that DTIC should offer training courses on how to use the system once it exists. The training could be a combination of computer based training, webinars or in person, instructor led courses.

Other groups indicated that providing analytical tools that assist in processing the data is another useful supporting function. The tools of interest to this particular group include data visualization tools that can handle large datasets as well as file comparison tools that can compare datasets to determine if the data is the same or the different. The specific comments made by the stakeholders are listed in Table 3.

Table 3: Stakeholder Comments - Supporting Activities

Ref#	Stakeholder Comment	Stakeholder
SA-1	DTIC may consider exploring storing open source software; however Government rights issues and intellectual property rights may be difficult to handle.	AFRL
SA-2	DTIC should have the capabilities of storing data, particularly where data may be orphaned or the collecting activity wants to share the data, but does not want to host the data.	AFRL
SA-3	The community would benefit from having data visualization tools included as part of a software library.	ARL
SA-4	The community would benefit from having data comparison tools that are capable of identifying similarities/differences between datasets.	ARL
SA-5	DTIC should define data standards.	NRL
SA-6	DTIC should create guidebooks for collecting/documenting data.	NRL
SA-7	DTIC should provide training on using the system.	NRL
SA-8	DTIC should consider implement an Electronic Notebook for researchers to capture data during their research, prior to producing a final document deliverable. The notebook would allow users to capture all the relevant information needed to describe the measurement data (by providing templates) as to what information, and how to format it. This will ensure that the information is captured during the project ad eliminates the needs to go back later to gather collect it.	NRL

3.1.4 Capabilities

The stakeholders expressed a variety of views on the capabilities needed within the MSDD. These comments covered a gamut of areas. The general consensus is that the MSDD needs to have provisions for storing data at DTIC, or a DTIC contracted site, to store datasets that are either orphaned or that the data owner unable to host on their own. The storage requirements, particular for “large datasets” can be substantial, potentially requiring petabytes of storage. The datasets may be single files, collections of files, or databases.

The ability to rate a datasets is another feature that was discussed. Overall, it was agreed that some type of rating system needs to be in place. Today, the data set quality is primarily determined by the experts in the different Community of Interests (COI). People not familiar with the data currently obtain an expert's advice on the data. Some suggestions are that a type of social networking capability will allow users to exchange this type of information.

In terms of access, it is clear that there will eventually need to be unclassified and classified versions of the MDSS. It is also important to allow university researchers to access the data.

The ability to link the datasets with supporting information, such as technical reports, is an important feature as well. This can present challenges as the data and technical reports may not be delivered at the same time. Furthermore, the technical reports will nearly always reside at DTIC which is not necessarily the case with the data. The links will need to be periodically checked to ensure they are still valid.

Developing a good set of metadata will be a key in establishing the success of the system. The metadata must include general information for searching the data, an equivalent of an electronic SF-298 form, as well as specific information regarding the actual measurement data. The latter will be domain specific. The comments regarding the datasets are shown in Table 4.

Table 4: Stakeholder Comments - Dataset Service Capabilities

Ref#	Stakeholder Comment	Stakeholder
CAP-1	A social networking mechanism is one method for allowing users to rate the quality of datasets (AFRL). There is value in a peer review of data.	AFRL
CAP-2	The data originator should be able to acquire information on the data requestors.	AFRL
CAP-3	The metadata should include a "dispose of" date.	AFRL
CAP-4	Large datasets can range in size from 25 to 100 TB.	AFRL
CAP-5	Research papers (reports) that pertain to each data set should be tracked and a link established that relates the two.	AFRL
CAP-6	The MSDD needs to support multiple levels of classification.	AFRL
CAP-7	The MSDD should focus on smaller datasets (not just focus on large datasets) at the outset. The focus can shift to large datasets later.	AFRL
CAP-8	DTIC should host/store DoD datasets; not just provide metadata linkage to the datasets. A significant number of data providers do not want to take on the administrative cost of providing the data to others (e.g., getting permission to host the data, managing the hardware/network issues, getting/maintaining the security certifications, and distributing the data).	AFRL
CAP-9	DTIC should attend and promote the dataset effort at relevant sensor/community-of-practice conferences (e.g., Tri-service Radar Symposium) to identify and collect new datasets.	AFRL

Ref#	Stakeholder Comment	Stakeholder
CAP-10	Processed data needs to have a lot of metadata to explain the contents, processing techniques, parameters, etc.	AFRL, NRL
CAP-11	DTIC should offer data standards to aid organizations in creating standard datasets.	ARL
CAP-12	DTIC needs to provide protected, authenticated dataset services that include a data vetting process.	NRL
CAP-13	Data collected by universities needs to be included.	NRL
CAP-14	The MSDD should contain documented “blind datasets” (i.e., ground truth is only known by a few people) is useful for supporting testing efforts. The ground truth is not made available to the general user community.	OSD

3.1.5 Challenges

There are a number of challenges associated with the MSDD project. The challenges cover a wide range of issues. For example, the releasability of the datasets presents several challenges. Traditionally, the data owner takes responsibility for the distribution of the information, generally making decisions on a case-by-case basis. The proposed program changes that paradigm for data owners that choose to participate. In some cases, the data owners may welcome the change; however, others are anticipated to decline to participate.

The issue of the scope of the distribution of the datasets was discussed. Ideally, the MSDD will be able to support data requests from not only Government agencies and DoD contractors, but also universities, members of the five English speaking nations, and NATO members.

The data quality issue needs to be addressed from several aspects. The first is the actual quality of the measurements and metadata as previously discussed. Traditionally, the COI members have developed trust relationships based upon personal experience with the data provider. Some existing data distribution sites, such as NASAs Data Explorer, allow users to search for datasets provided by specific individuals. Several stakeholders commented on the need to have some method of rating the datasets. This can be accomplished by several methods including individual user feedback or a COI feedback. The second part of the data quality issue pertains to possible degradations due to processing of the data (where only processed data is available, as opposed to the original raw data), as well any errors produced by using data compression algorithms that maybe applied to reduce the storage space.

The stakeholder comments on the challenges are provided in Table 5.

Table 5: Stakeholder Comments - Challenges

Ref#	Stakeholder Comment	Stakeholder
CH-1	The administration (registering users, validating, determining permissions) process is a large obstacle. There can be a significant cost to maintaining and administering datasets.	AFRL
CH-2	The determination of the releasibility of the datasets is a problem.	AFRL
CH-3	The MDSS would ideally be accessible to NATO members, Australia and New Zealand (the remaining Five Eyes countries).	AFRL
CH-4	There are a number of Information Assurance issues that need to be addressed.	AFRL
CH-5	Datasets should be available at no cost to the consumer, except if there is a cost for media and labor to copy large datasets.	AFRL
CH-6	Data ownership may be a problem. At NRL, each division owns their data and they determine the distribution.	NRL
CH-7	Researchers collaborate with Universities. Universities may need data or may be able to host data under contract to the Government.	NRL
CH-8	Should universities be allowed to host data?	OSD
CH-9	Some caution needs to be exercised on what is an available asset (some hardware/software may not actually be available or may have been decommissioned).	OSD
CH-10	Data/Information quality issues may arise as data is compressed or processed. A definition of the quality needs to be established.	OSD

3.1.6 Datasets

The stakeholders suggested several specific and some generic datasets that they either have, or believe exist within their organizations. The stakeholders also recommended contacting individual COI's to gather additional information on the types of data available. A list of these datasets is provided in Table 6. In addition, OSD provide a listing of datasets that they are considering as part of the JDMS. This list is not included in this report because it is sensitive to the JDMS program.

Table 6: Stakeholder Comment - Datasets

Ref#	Stakeholder Comment	Stakeholder
DA-1	Datasets available from AFRL/RI: a. "Swathbuckler" (Radar) dataset generated as part of a The Technology Cooperation Program (TTCP) experiment. b. Fusion data c. Exploitation data d. Cyber/network traffic e. AFRL SITES data (from Newport, etc.)	AFRL

Ref#	Stakeholder Comment	Stakeholder
DA-2	Data available from the ARL includes: a. PCAP (packet capture) network packets, particularly associated with mobile and wireless networks; b. Intrusion detection system data; c. Meteorological data; and d. Chemical and biological data.	ARL
DA-3	From various community of interests, such as sensor data: a. Acoustics b. Oceanographic (includes Fleet Numerical Meteorology and Oceanography Center, NOAA, NGA, Naval Ocean and Atmospheric Research Lab - St. Louis, and TRMC) c. Related University Data	NRL
DA-4	Clementine dataset from NRL	NRL

3.2 Metadata Study Results

A number of ongoing efforts are establishing and investigating standards in terms of metadata, data classification and data sharing. The combinations of these efforts can provide guidance for what the MSDD automated search strategy should/should not include and define the kind of data that the DTIC is looking for. In terms of taxonomy some ongoing efforts, such as the National Information Exchange Model, have already answered this question by using eXtensible Markup Language (XML) combined with Universal Core (UCore) practices.

The metadata elements that are to be used for DoD related data exchange and labeling projects should be the ones defined in the DoD Metadata registry v7.2+. From the ongoing efforts, the notable mentions relevant to this effort include:

1. DoD Metadata Registry v7.2+ and Clearinghouse
2. Department of Defense Discovery Metadata Specification (DDMS)
3. Defense Knowledge Online (DKO)
4. The Open Archives Initiative Protocol for Metadata Harvesting (OIA)
5. UCORE
6. NIEM (National Information Exchange Model) and LEMS (Logical Entity Exchange Specification)
7. Metadata Tools for Geospatial Data

3.2.1 DoD Metadata Registry (MDR) v7.2+ and Clearinghouse[2]

The DoD Metadata Registry contains numerous XML Schema documents defining data elements from across the Department. XML Schema developers and data modelers are typically interested in reusing some of the entities defined in the XML Schemas registered on the DoD MDR rather than re-creating their own. ISO/IEC 11179 (formally known as the ISO/IEC 11179 Metadata Registry (MDR) standard) is an international standard for representing metadata for an organization in a Metadata Registry.

The DoD Metadata Working Group (MWG) consists of members of the Defense Information Systems Agency (DISA) Engineering Staff, Namespace Managers, and representatives of related Working Groups, members of the MDR Operations Staff, DoD Developers, and other interested parties. The DoD MWG is responsible for ensuring that the DoD Metadata Registry and Clearinghouse (and other metadata management capabilities) meets the goals of net-centricity and Enterprise metadata requirements.

The Metadata Registry Architecture is shown in Figure 4.

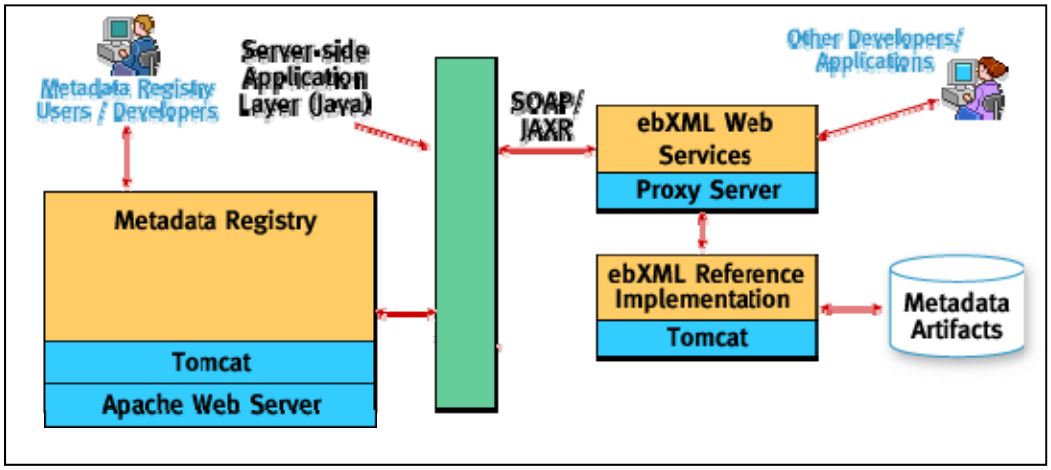


Figure 4: Metadata Registry Architecture

3.2.1.1 Department of Defense Discovery Metadata Specification (DDMS)

The DoD Discovery Metadata Specification (DDMS) defines discovery metadata elements for resources posted to community and organizational shared spaces.

"Discovery" is the ability to locate data assets through a consistent and flexible search. Visibility, accessibility, and understandability are the high priority goals of the DoD Net-Centric Data Strategy. DDMS specifies a set of information fields that are to be used to describe any data or service asset, i.e., resource, that is to be made discoverable to the Enterprise, and it serves as a

[2] https://metadata.dod.mil/mdr/help.htm?page=faqs#urn:uuid:fee4e3fc-519a-404f-b42c-86a510085818

reference for developers, architects, and engineers by laying a foundation for Discovery Services. According to the working group of DDMS, the DDMS will be employed consistently across the DoD's disciplines, domains and data formats. The following are relevant links to the information source and standard itself:

http://metadata.dod.mil/mdr/irs/DDMS/
http://metadata.dod.mil/mdr/ns/DDMS/2.0/DDMS_v2.0.zip
http://metadata.dod.mil/mdr/ns/DDMS/current/

The DDMS System Architecture is shown in Figure 5.

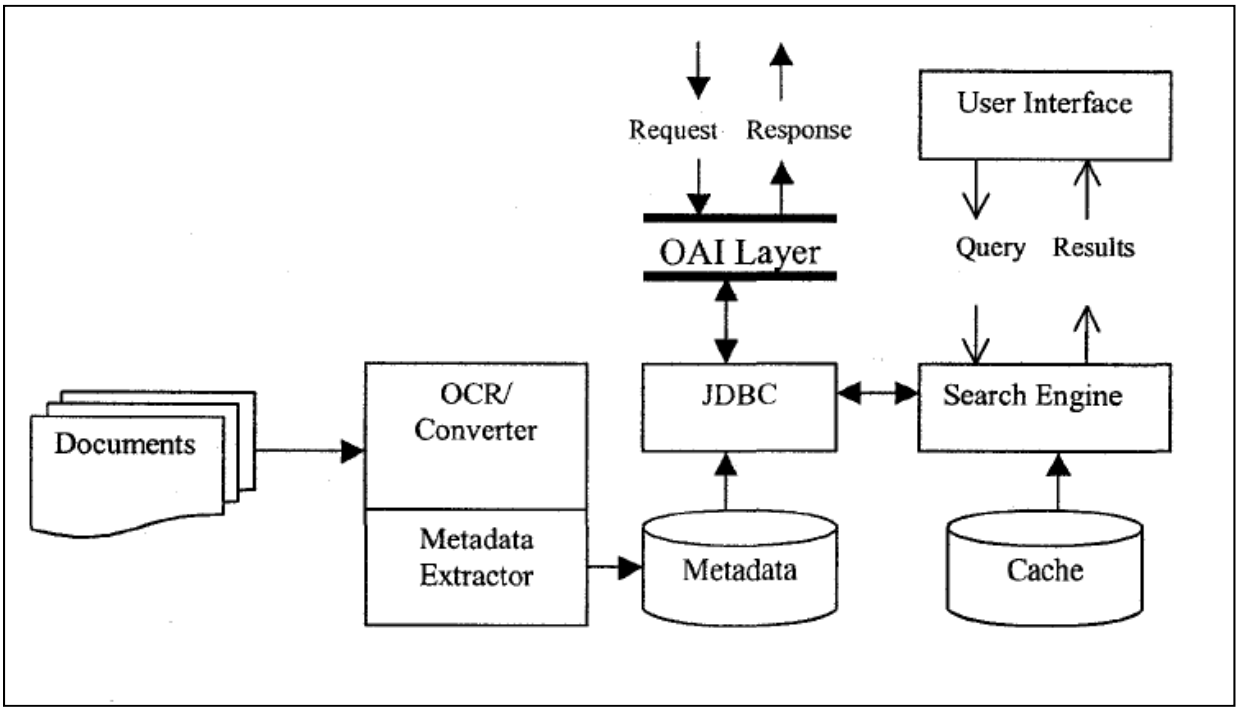


Figure 5: DDMS System Architecture

Table 7 depicts the category sets defined by DDMS.

Table 7: DDMS Category Overview

DDMS Category Sets: Introduction and Definitions			
Security	**Resource**	**Summary Content**	**Format**
security	title subtitle	subject categoryQualifier categoryCode categoryLabel keyword	format mediaFormat extentQualifier extent medium
	creator publisher contributor pointOfContact person name surname userID organization phoneNumber emailAddress organization name phoneNumber emailAddress webService name phoneNumber emailAddress	geospatialCoverage geographicIdentifier geographicBoundingBox geographicBoundingGeometry postalAddress verticalExtent facilityBENumber facilityOsuffix region name westboundLongitude eastboundLongitude northboundLatitude southboundLatitude polygon point street city state postalCode countryCodeQualifier countryCode province minimumVerticalExtent maximumVerticalExtent	
	identifier qualifier value	temporalCoverage dateStart dateEnd timePeriod	
	date created posted validTil infoCutOff	virtualCoverage virtualAddress networkProtocol	
	rights privacyAct intellectualPropertyRights copyright	description	

DDMS Category Sets: Introduction and Definitions			
Security	**Resource**	**Summary Content**	**Format**
	language qualifier value	Related Resources relationship direction RelatedResource qualifier value link href label title role type	
	type qualifier value		
	source qualifier value schemaQualifier schemaHref		

3.2.1.2 Defense Knowledge Online (DKO)

The DoD requires an information sharing environment that supports secure access to disparate, cross-service capabilities and information as an enterprise collaborative environment for war fighting, business, and intelligence users.

The focus of Defense Knowledge Online (DKO) is the requirement for a DoD-wide collaborative enterprise. DKO is currently hosted as part of the Army Knowledge Online (AKO) portal. The Defense Information Systems Agency (DISA) Net-Centric Enterprise Services (NCES) Program Management Office is responsible for providing a suite of core enterprise services (CES) that improves the ability to collaborate and discover/subscribe to existing information sources. NCES is required to facilitate access to these capabilities and provide them from a web browser or through existing Command/Service/Agency (C/S/A) Portals.

3.2.1.3 The Open Archives Initiative Protocol for Metadata Harvesting[3]

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on *metadata harvesting*. The Open Archives Initiative is an organization formed by a broad range of librarians, publishers, researchers, and archivists. Its goal is to create simple standards to support interoperability among heterogeneous digital libraries. The OAI-PMH provides an application-independent interoperability framework.

OAI-PMH requests are expressed as Hypertext Transfer Protocol (HTTP) requests such as HTTP GET or POST methods. All responses to OAI-PMH requests must be well formed extensible Markup Language (XML) instance documents. The returned XML record has three parts:

- Header – information common to all records and necessary for the harvesting process,
- Metadata – metadata elements of returned records, and
- About – optional container to hold data about the metadata part of the record

Service providers harvest metadata from data providers using the OAI-PMH and use the harvested metadata as the basis for building value-added services. These archives would then act as a federation of repositories, by indexing documents in a standardized way so that multiple collections could be searched as though they form a single. This service is called cross-archive search. While current Web search engines usually deal with semi-structured data, cross-archive search engines using the OAI-PMH framework should exploit structured metadata describing the core information properties.

[3] http://www.openarchives.org/OAI/openarchivesprotocol.html

3.2.1.4 Universal Core (UCore)

Universal Core (UCore) has been approved for incorporation into the Department of Homeland Security (DHS) Enterprise Architecture Technical Reference Model and will be included in the pending update to the DoD Manual for Implementing Net Centric Data Sharing (DOD 8320.02M).

UCore is a federal initiative that supports the National Information Sharing Strategy and all associated Departmental / Agency strategies. UCore enables information sharing by defining an implementable specification (XML Schema) containing agreed upon representations for the most commonly shared and universally understood concepts of Who, What, When, and Where.

UCore is designed to be simple to understand, explain, and implement. It is small, containing a minimal set of objects with broad applicability across a wide range of domains. UCore is built on an extensible framework that permits users to create more detailed exchanges tailored to their mission or business requirements. The specification is based on, and leverages, existing commercial standards, governmental standards, and best practices. The UCore validation processes and tools provide a means to consistently achieve definable levels interoperability, promoting machine understanding between both anticipated and unanticipated users.

Technically, UCore is an eXtensible Markup Language (XML) based information exchange specification and implementation profile. It provides a framework for sharing the most commonly used data concepts of who, what, when, and where and serves as a starting point for interagency information sharing and data level interoperability. It also provides the framework, metadata, extension rules, security markings, and physical schema to permit content to be exchanged between heterogeneous IT infrastructures.

Because UCore has been designed to be interoperable with NIEM and LEXS, current NIEM-based systems will not need to deviate from existing implementations to share information via UCore. The NIEM program is fully committed to ensuring that future versions of NIEM and LEXS will be similarly compatible with UCore.

3.2.1.5 NIEM (National Information Exchange Model) and LEXS (Logical Entity Exchange Specification)[4]

The National Information Exchange Model (NIEM) is a partnership of the U.S. Department of Justice (DOJ) and DHS. It is designed to develop, disseminate and support enterprise-wide information exchange standards and processes that can enable jurisdictions to effectively share critical information in emergency situations, as well as support the day-to-day operations of

[4] http://www.niem.gov/ and http://www.lexs.gov/

agencies throughout the nation. It leverages the data exchange standards efforts successfully implemented by the Global Justice Information Sharing Initiative (Global) and extends the Global Justice XML Data Model (GJXDM) to facilitate timely, secure information sharing across the whole of the justice, public safety, emergency and disaster management, intelligence, and homeland security domains. NIEM has 10 domains:

1. Chemical, Biological, Radiological, Nuclear (CBRN)
2. Emergency Management
3. Immigration
4. Infrastructure Protection
5. Intelligence
6. International Trade
7. Justice
8. Maritime
9. Screening
10. (Youth and) Family Services

NIEM also has some readily available tools such as the graphical data model browser. Since it is written in Java, we assumed that it could be utilized in any platform. Figure 6 is a screenshot of the tool.

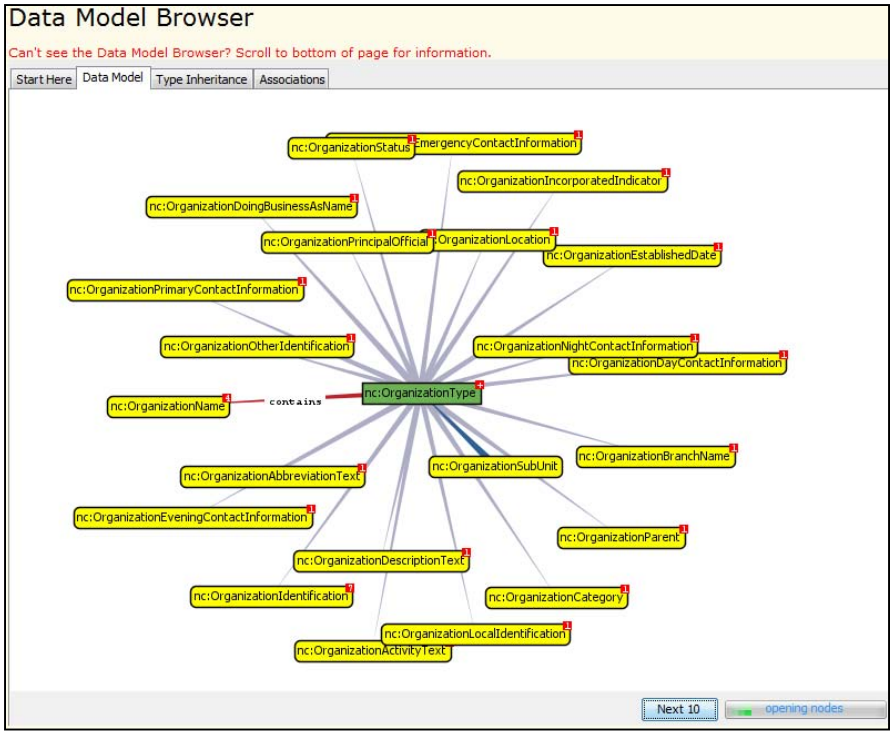


Figure 6: Graphical Data Model Browser

LEXS is a comprehensive, NIEM-based, framework for the development of information exchanges. Initially developed for the law enforcement information sharing program at US Department of Justice, LEXS is now being widely used in criminal justice community at large, as well as by the homeland security, intelligence and other communities.

3.2.2 Sample Datasets

A number of datasets were identified as part of the effort. The sample datasets were used to perform a sampling of the types of data available, the locations, and if XML/XSD exists for the data. A list of the sample data types investigated is provided in Table 8.

Table 8: Sample Datasets

Dataset Title	Subsets	URL	XML/XSD	Metadata
14th Weather Squadron	Aircraft Observations Agrmet / Agrimet Cloud/Merged Analysis- Real Time Nephanalysis (RTNEPH) GHCN Precipitation GPCP Precipitation Joint Lightning Legates Lightning Snow Analysis Snow Climatology Summary Of The Day Surface Weather Observations Upper Air Analysis - (GFS) Upper Air Observations Weather Information Network Display System (WINDS)-Kennedy Space Center, FL Weather Information Network Display System (WINDS)-Vandenberg, CA Wind-Stratified Conditional Climatology (WSCC)	https://notus2.afccc.af.mil/SCISPublic/services/databases.asp	N/A	Partial Geospatial Metadata
Air Force Weather information	Satellite Imagery Radar Imagery Worldwide Local Weather Local Weather Map Meteorological Information	http://preview.afnews.af.mil/afwa/weatherproducts/index.asp	N/A	N/A
The Navy's oceanographic portal	Meteorology Products Oceanography Products Tropical Applications Climatology and Archived Data	http://www.usno.navy.mil/FNMOC/meteorology-products-1	N/A	N/A
Global Ocean Data Assimilation Experiment (GODAE)	Argo USGODAE GDAC AATSR (Advanced Along Track Scanning Radiometer) COAMPS - Coupled Ocean Atmosphere Mesoscale Prediction System	http://www.usgodae.org/index.html	N/A	TESAC, GTS stream

Dataset Title	Subsets	URL	XML/XSD	Metadata
	NOGAPS -Navy Operational Global Atmospheric Prediction System NOGAPS ANGM - Navy Operational Global Atmospheric Prediction System Angular Momentum NOGAPS Computational Grids FNMOC High Resolution SST/Sea Ice Analysis for GHRSST FNMOC High Resolution Ocean Analysis for GODAE GOES - FNMOC GOES 10 Satellite Retrievals MCSST - AVHRR SST retrievals from FNMOC Ocean QC Process Meteorological Data PROFILE - Fixed/drifting buoy, bathy, and PALACE float data SFCOBS - Surface Observations: Ship, fixed/drifting buoy, and CMAN in-situ surface temperatures SWH - FNMOC Sea Wave Height from Satellite Altimeters TRACK - FNMOC Ship Track SST measurements WW3 - FNMOC Wave Watch III WW3 Mediterranean - FNMOC Wave Watch III TROPICAL CYCLONE - JTWC/NHC Tropical Cyclone Warnings LAS GODAE Modelers Output Site World Ocean Atlas 2001 World Ocean Atlas 1998 Navy GDEM Climatology NAVO ERS-2 SSH NAVO GFO SSH NAVO JASON-2 SSH NAVO TOPEX SSH NAVO GOES SST NAVO LAC SST NAVO MCSST - NAVOCEANO daily sea surface temperature (SST) retrievals. US Navy 5 Minute Bathymetry DBDBV University of Washington Applied Physics Lab Seaglider AUV GFS - NOAA NCEP GFS Model Smith and Sandwell Satellite Bathymetry			

Dataset Title	Subsets	URL	XML/XSD	Metadata
	NON-OPERATIONAL JASON-1 SSH AMSR-E (Advanced Microwave Scanning Radiometer) SAF (Meteo France Satellite Application Facility Mirror) USGS ETOPO5			
1998 DARPA Intrusion Detection Evaluation Data Set	There were two parts to the 1998 DARPA Intrusion Detection Evaluation: an off-line evaluation and a real-time evaluation.	http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/1998data.html	N/A	N/A
Universal Core Community	Integrated Weapons of Mass Destruction Toolset Objective Gateway Common Data Framework Radio Frequency Propagation Service Strategic Knowledge Integration Web (SKIWeb) UCore and NIEM Tagging Proof of Concept (POC) UCore-Semantic Layer (UCore-SL)	https://www.UCore.gov/UCore/	Both	Yes XML format
US Government XML Working Group	Standards & Guidelines Registries/Repositories	http://www.xml.gov/	Both	XML, various
Department of Defense's transformation initiative	Video content	http://www.defenselink.mil/transformation/images/video	N/A	N/A
eCoastal is a geodatabase structure that includes coastal related, SDS (Spatial Data Standard) compliant datasets (Bathymetric and condition surveys, dredging information, NOAA Charts	Coastal data	http://ecoastal.usace.army.mil/faq.asp	XML data	DBMS dependent
Metadata Tools for Geospatial Data	This page leads to summaries of most of the known metadata tools used for documenting geospatial data and serving geospatial metadata. It includes tools for entering and editing metadata and utilities for preprocessing, extracting, post processing, validating, and viewing metadata. Most of these tools were designed to help complete Content Standards for Digital Geospatial Metadata (CSDGM) metadata, but several have been tuned to produce specific local metadata profiles.	http://www.sco.wisc.edu/wisclinc/metatool/ http://www.fgdc.gov/metadata/iso-metadata-editor-review	N/A	Some

3.2.3 Commonly used Metadata Standards

The following metadata schemas are selected based on their applicability to the university community. The standard might be of importance when considering strategies in future MSDD investigations.

3.2.3.1 Data Documentation Initiative (DDI)

The DDI is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioral sciences.

3.2.3.2 Dublin Core

The Dublin Core is a flexible 15-element metadata set. The Dublin Core is used by organizations such as libraries and government agencies for text, images, and other resources.

3.2.3.3 Encoded Archival Description (EAD)

EAD is a data structure standard for encoding archival finding aids, developed for use by the archivists and manuscript curators.

3.2.3.4 Federal Geographic Data Committee (FGDC)

The FGDC standard was created to provide a common set of terminology and definitions for the documentation of digital geospatial data. Without the essential information provided in FGDC metadata many spatial data files would not be considered reliable data sources.

3.2.3.5 Instructional Management Systems (IMS)

The IMS Learning Resource Meta-data Information Model identifies a subset of IEEE LOM meta-data elements to be used to describe learning materials in various types of learning systems.

3.2.3.6 Metadata Encoding and Transmission Standard (METS)

METS provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users). Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model.

3.2.3.7 ONline Information eXchange (ONIX)

ONIX is an international standard for representing book, serial, and video product information in electronic form. Many on-line book traders such as Amazon and Barnes & Noble use this metadata standard to transfer information about their products.

3.2.3.8 Sharable Content Object Reference Model (SCORM)

SCORM uses the IEEE LOM element set for descriptive metadata, and includes guidelines on the XML packaging of the metadata. SCORM draws on a variety of standards to create reference model specifically for learning objects.

3.2.3.9 TEI (Text Encoding Initiative)

An encoding standard for textual documents used to describe the physical and logical structure of textual material for the purpose of research analysis and data interchange. A header containing bibliographic information and provenance precedes the full encoding.

3.2.3.10 Visual Resources Association (VRA)

A core element set used to create records to describe works of visual culture as well as the images that document them. Used by image archives in museums and libraries.

4.0 Conclusions

This effort focused on performing two tasks in support of the MSDD, namely surveying representatives from the stakeholder community and performing an initial study on metadata used to describe datasets. The key findings of this effort are summarized below:

- The stakeholders support the MSDD concept from a technical perspective and they indicated that DTIC is the right agency to lead the effort.
- The stakeholders stated that a business case must be established to prove that there are financial merits to establishing the capability.
- DTIC needs to have the capability of storing data as well as leverage third parties that host datasets.
- DTIC should provide guidance on the metadata for the MSDD.
- DTIC should provide an access control service for third party sites to authenticate and authorizer data distribution to users.
- DTIC should leverage and/or coordinate with other ongoing related projects.

The results of the effort indicate that MSDD will benefit the DoD RDT&E community and warrants additional investigation and a move towards implementing the capability by leveraging existing DTIC assets. The specific recommendations for continuing the effort are described in the next section.

5.0 Recommendations

The results of the study serve as grounds for the recommendations discussed in the following sections.

5.1 Recommendation 1: Develop a Detailed Program Plan

The Managed Services effort would benefit from a detailed program plan. An initial plan was provided as part of the proposal for the initial effort. The plan should be revised to incorporate the additional steps required to start an add-on capability to DTIC's existing infrastructure. This includes establishing a small pilot program to investigate activities such as:

- Identifying additional datasets and communities of interest
- Identify key challenges and perform trade studies on potential solutions, such as the Shared Access Control prototype.
- Requirements definition
- Investigate data and metadata ontologies
- Demonstrate the business and financial advantages of the program.

5.2 Recommendation 2: Acquire Additional Stakeholder Feedback

The stakeholders interviewed as part of this effort represent a relatively small sample of the RDT&E community. Each group recommended addition organizations or individuals that they believed may be interested in the program, as either a data source, data provider, or have technical or managerial insight into related programs and technologies. It is important to continue to involve additional stakeholders, either as part of the proposed demonstration (refer to Recommendation 4), or with individual/group meetings. Table 9 lists additional potential stakeholders identified during the interviews.

Table 9: Additional Stakeholders Recommended by the Community

Organization/Individual	Suggested by
AFRL Executive Director, who oversees the entire AFRL S&T program.	AFRL
Networking and Information Technology Research and Development (NITRD) Program	AFRL
TARDEC/TACOM – Tank Automotive Research, Development and Engineering Center	ARL
ARDEC – Armament Research, Development and Engineering Center (Picatinny)	ARL
CERDEC – Communications-Electronics Research, Development, and Engineering Center (Ft. Monmouth)	ARL

Organization/Individual	Suggested by
STTC – Simulation & Training Technology Center (Orlando)	ARL
AMRDEC – Aviation and Missile Research, Development and Engineering Center (Huntsville)	ARL
ECDC – Edgewood Chemical Biological Center	ARL
NSRDEC – Natick Soldier Research, Development and Engineering Center (Natick, MA)	ARL
OPTEC - Operational Test and Evaluation Command	ARL
Network Science Collaborative Technology Alliance (Includes University of Pennsylvania, University of Illinois, and others)	ARL
HPC Community – a list of specific individuals as provided.	HPCMP
TRMC – DoD Test resource Management Center	NRL
JITC – Joint Interoperability Test Command	NRL
Library of Congress	NRL

The recommendation is that a limited number of these stakeholders be interviewed to gather additional perspective.

The stakeholder interaction also seeks to identify other similar projects, such as JDMS, that can be a leveraging opportunity. An additional approach is to present the program at a conference or hold a MSDD workshop after a major conference.

5.3 Recommendation 3: Solicit Additional Requirements

The initial effort identified a few features and goals to assist in defining the overall program but does not provide sufficient detail to develop a detailed requirements specification. The exception is the distributed authentication and authorization approach that was discussed with DTIC and AFRL representatives. The recommendation is to develop a complete requirement set that can be used to support a full development effort.

5.4 Recommendation 4: Develop a Shared Access Control Prototype

As a result of discussions with AFRL, a white paper was prepared that proposes a prototype to demonstrate the use of a shared access control system that would allow third parties to leverage the DTIC Online Access Control system. The third parties represent organizations that function as data stores. One of the challenges faced by these organizations is the cost and complexity of administrating and maintaining the user lists. In particular, trying to register, validate, and determining the authorization parameters and rules is a daunting task. DTIC has solved a number of these problems as part of their technical document distribution capability. The white paper

proposes leveraging these capabilities and providing the third party sites with a “yes/no” result in response to specific access requests. This paper has been provided under a separate cover.

5.5 Recommendation 5: Develop a Data Search Prototype

A second prototype would focus on examining methods for searching for datasets within the DoD community. If this initiative moves forward, the initial concentration will focus on datasets offered by the community. The next stage is to identify datasets that are available online, but have not been submitted for inclusion into the effort. This effort will seek to identify the datasets and their owners using a combination of web crawling and web service discovery approaches. This approach has merits to increase the collection size as the program matures. In an operational scenario, the data owner would be contacted and offered the option of participating in the program.

6.0 Glossary

The acronyms and terms used in this report are defined in Table 10.

Table 10: Acronyms and Terms

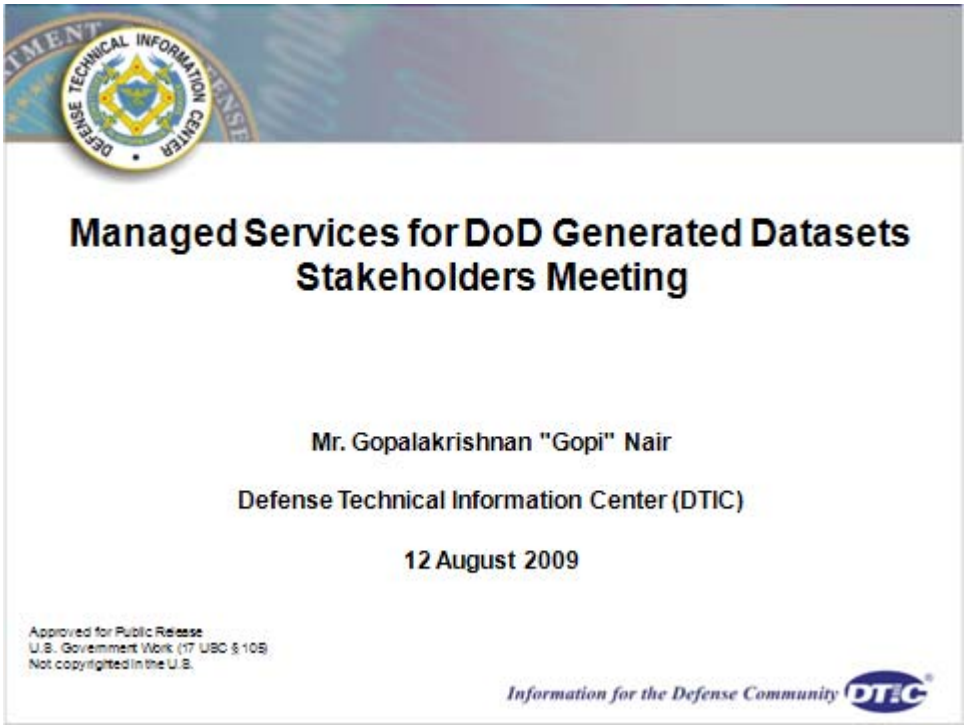
Term	Definition
AFRL	Air Force Research Laboratory
AKO	Army Knowledge Online
ARDEC	Armament Research, Development and Engineering Center
AMRDEC	Aviation and Missile Research, Development and Engineering Center (Huntsville)
ARL	Army Research Laboratory
CERDEC	Communications-Electronics Research, Development, and Engineering Center
CES	Core Enterprise Services
CBRN	Chemical, Biological, Radiological, Nuclear
C/S/A	Command/Service/Agency
DKO	Defense Knowledge Online
DISA	Defense Information Systems Agency
DHS	Department of Homeland Security
DoD	Department of Defense
DoJ	Department of Justice
DDMS	DoD Discovery Metadata Specification
DDI	Data Documentation Initiative
DTIC	Defense Technical Information Center
EAD	Encoded Archival Description
ECDC	Edgewood Chemical Biological Center
FGDC	Federal Geographic Data Committee
HPCMP	High Performance Computing Modernization Program
HTTP	Hypertext Transfer Protocol
IMS	Instructional Management Systems
JITC	Joint Interoperability Test Command
JDM	Joint Data Management
LEXS	Logical Entity Exchange Specification
METS	Metadata Encoding and Transmission Standard
MDR	Metadata Registry
MSDD	Managed Services for DoD-Generated Datasets
MWG	Metadata Working Group
NCES	Net-Centric Enterprise Services
NIEM	National Information Exchange Model

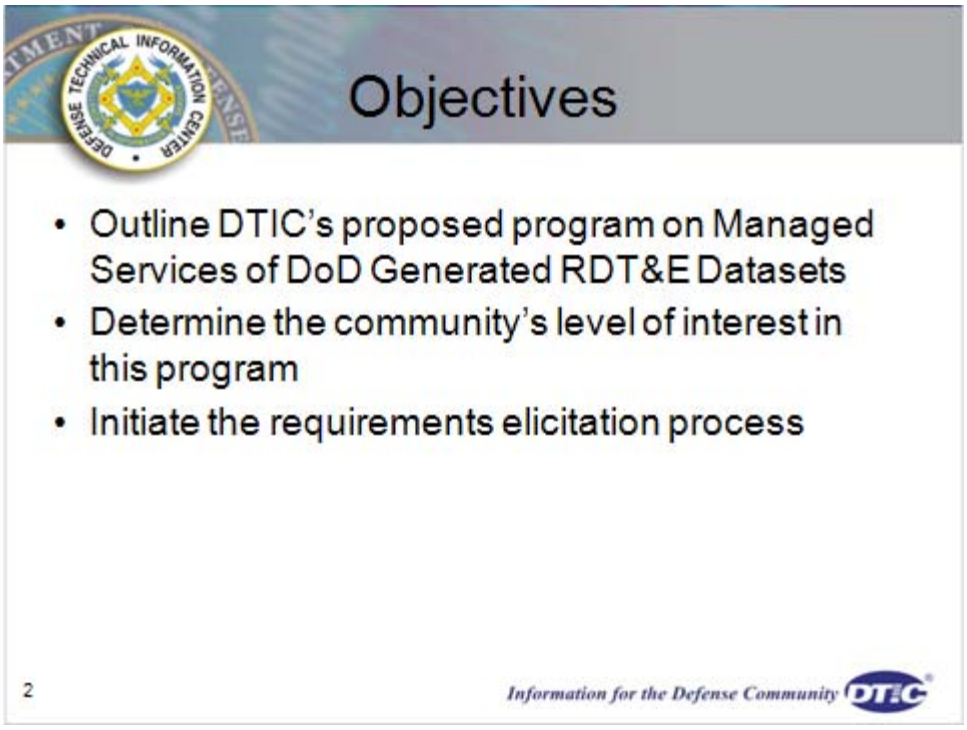
Term	Definition
NITRD	Networking and Information Technology Research and Development Program
NSRDEC	Natick Soldier Research, Development and Engineering Center
NRL	Naval Research Laboratory
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ONIX	ONline Information eXchange
OPTEC	Operational Test and Evaluation Command
OSD	Office of the Secretary of Defense
R&D	Research & Development
RDT&E	Research Development Test & Evaluation
SCORM	Sharable Content Object Reference Model
STTC	Simulation & Training Technology Center
TARDEC/TACOM	Tank Automotive Research, Development and Engineering Center
TEI	Text Encoding Initiative
TRMC	DoD Test resource Management Center
UCore	Universal Core
VRA	Visual Resources Association
XML	eXtensible Markup Language

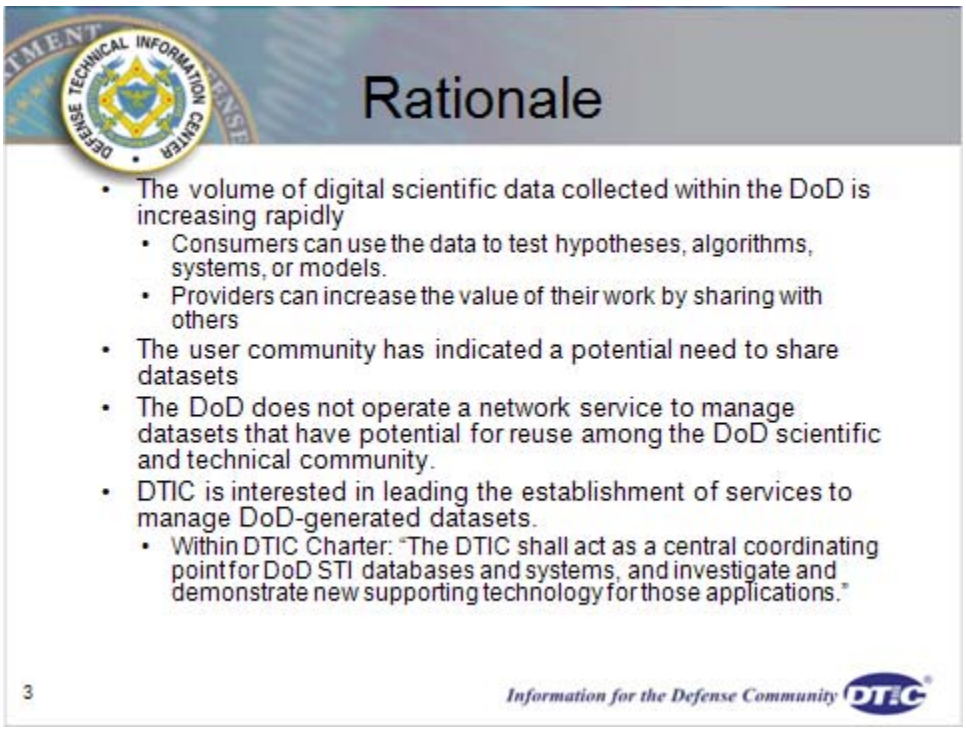
Appendix I

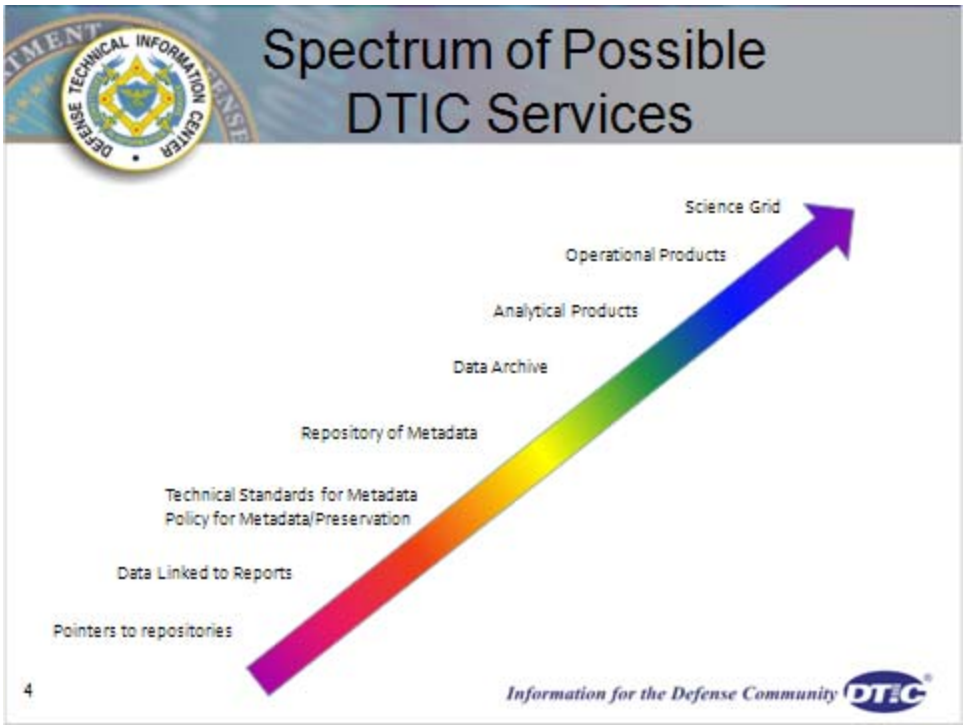
MSDD Stakeholders Meeting Presentation

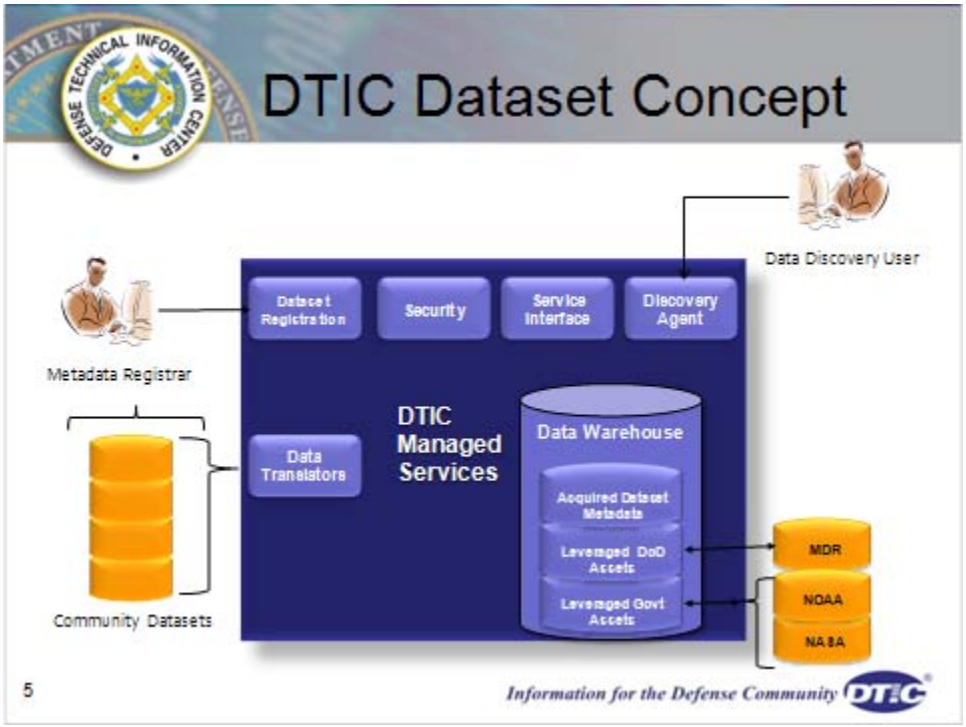
The purpose of the MSDD Stakeholders Meeting presentation developed for this effort was to gauge interest in DTIC providing Managed Dataset Services and identify initial requirements. The slides are shown on the following pages.

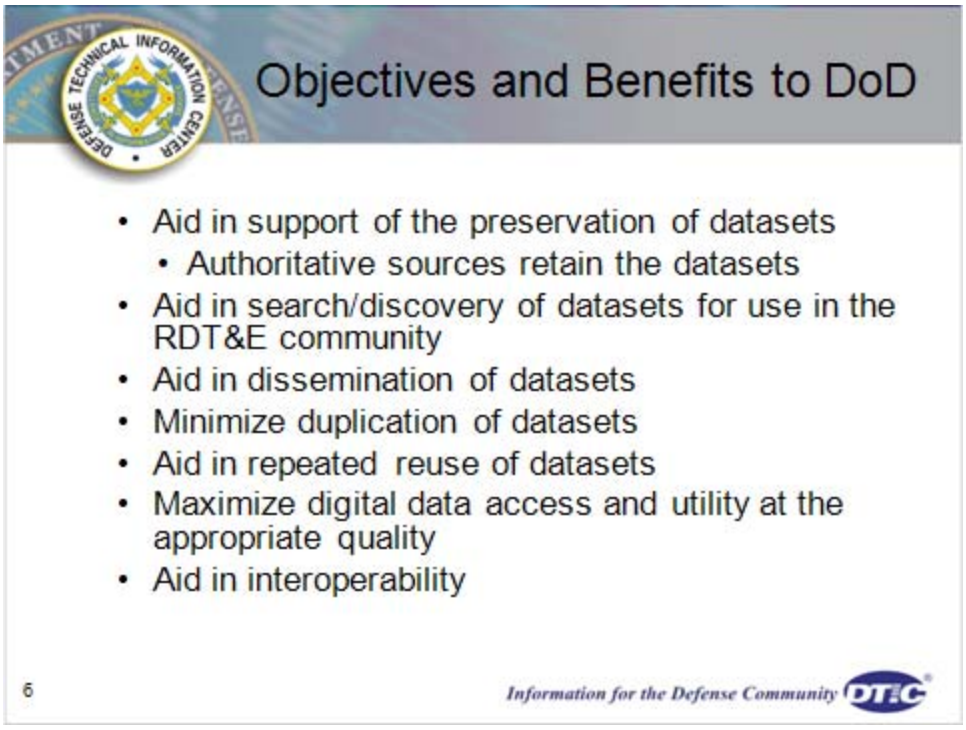


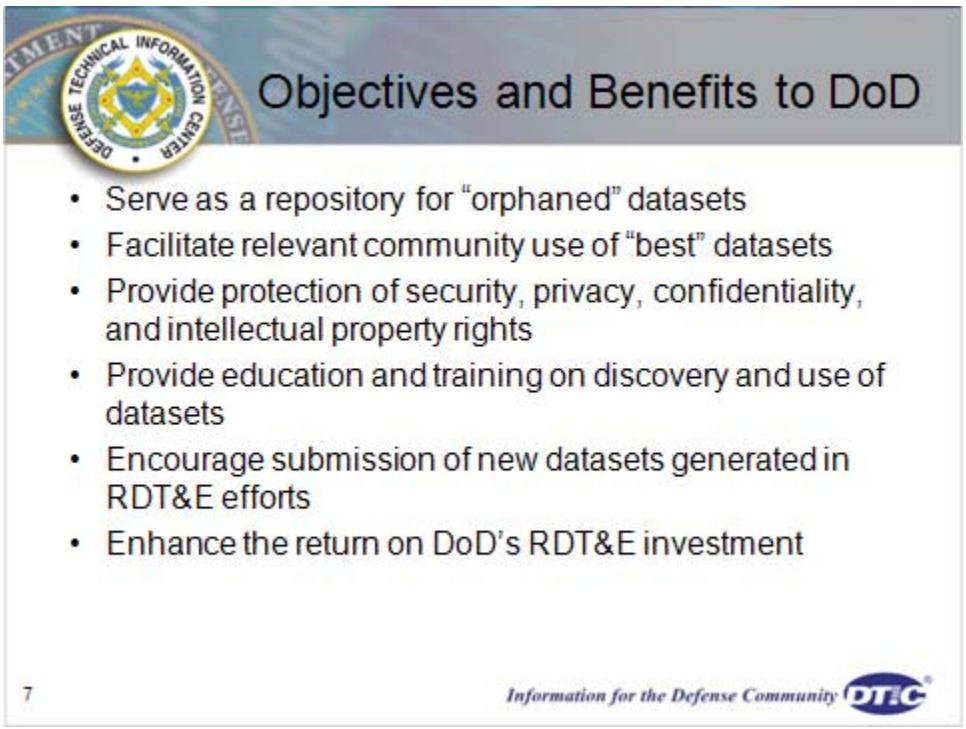

Objectives
• Outline DTIC's proposed program on Managed Services of DoD Generated RDT&E Datasets
• Determine the community's level of interest in this program
• Initiate the requirements elicitation process
2
Information for the Defense Community DTIC


Rationale
• The volume of digital scientific data collected within the DoD is increasing rapidly
• Consumers can use the data to test hypotheses, algorithms, systems, or models.
• Providers can increase the value of their work by sharing with others
• The user community has indicated a potential need to share datasets
• The DoD does not operate a network service to manage datasets that have potential for reuse among the DoD scientific and technical community.
• DTIC is interested in leading the establishment of services to manage DoD-generated datasets.
• Within DTIC Charter: "The DTIC shall act as a central coordinating point for DoD STI databases and systems, and investigate and demonstrate new supporting technology for those applications."
3
Information for the Defense Community DTIC


Spectrum of Possible DTIC Services
Science Grid
Operational Products
Analytical Products
Data Archive
Repository of Metadata
Technical Standards for Metadata
Policy for Metadata/Preservation
Data Linked to Reports
Pointers to repositories
4
Information for the Defense Community DTIC


DTIC Dataset Concept
Data Discovery User
Dataset Registration
Security
Service Interface
Discovery Agent
Metadata Registrar
DTIC Managed Services
Data Translators
Data Warehouse
Acquired Dataset Metadata
Leveraged DoD Assets
Leveraged Govt Assets
MDR
NOAA
NASA
Community Datasets
5
Information for the Defense Community DTIC


Objectives and Benefits to DoD
• Aid in support of the preservation of datasets
• Authoritative sources retain the datasets
• Aid in search/discovery of datasets for use in the RDT&E community
• Aid in dissemination of datasets
• Minimize duplication of datasets
• Aid in repeated reuse of datasets
• Maximize digital data access and utility at the appropriate quality
• Aid in interoperability
6
Information for the Defense Community DTIC


Objectives and Benefits to DoD
• Serve as a repository for "orphaned" datasets
• Facilitate relevant community use of "best" datasets
• Provide protection of security, privacy, confidentiality, and intellectual property rights
• Provide education and training on discovery and use of datasets
• Encourage submission of new datasets generated in RDT&E efforts
• Enhance the return on DoD's RDT&E investment
7
Information for the Defense Community DTIC

DTIC's Role in Digital Data Life Cycle

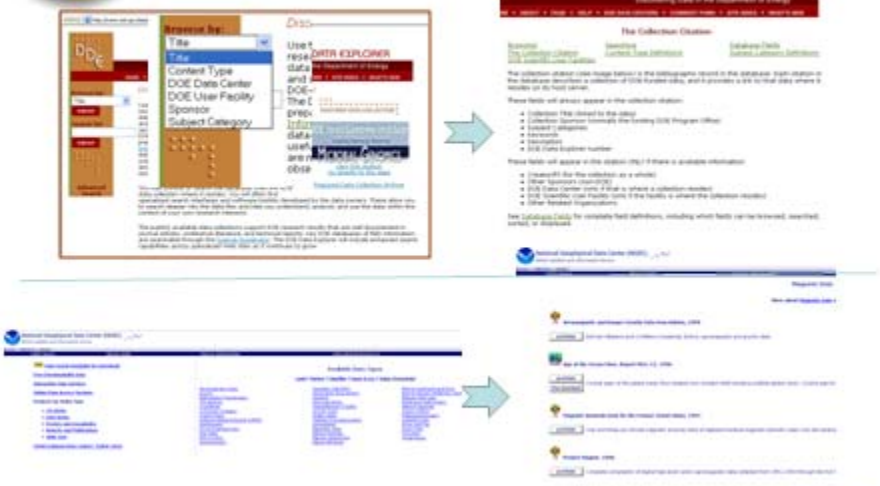
Life Cycle Phase	Role of DTIC Services for Data Preservation
Creation	Identify sources of data. Identify permanent location. Establish agreements and responsibilities
Ingestion or Acquisition	Establish technologies to access and search. Establish metadata. Provide visibility.
Documentation	Define metadata standards. Track usage and status
Organization	Develop repository standards. Conform to standards
Migration	Maintain awareness of dataset location and migration of data.
Protection	Implement quality control, access restrictions, user authorization, and trustworthiness
Access	Catalog and describe metadata. Disseminate information about available data. Support necessary methods for discovery. Provide community support
Disposition	Assume ownership of orphaned data. Remove unwanted information.

Based on "Harnessing the Power of Digital Data for Science and Society"
Report of the Interagency Working Group on Digital Data to the Committee of the Nation Science and Technology Council
January 2009 (http://www.nitrd.gov/about/Harnessing_Power.aspx)

8

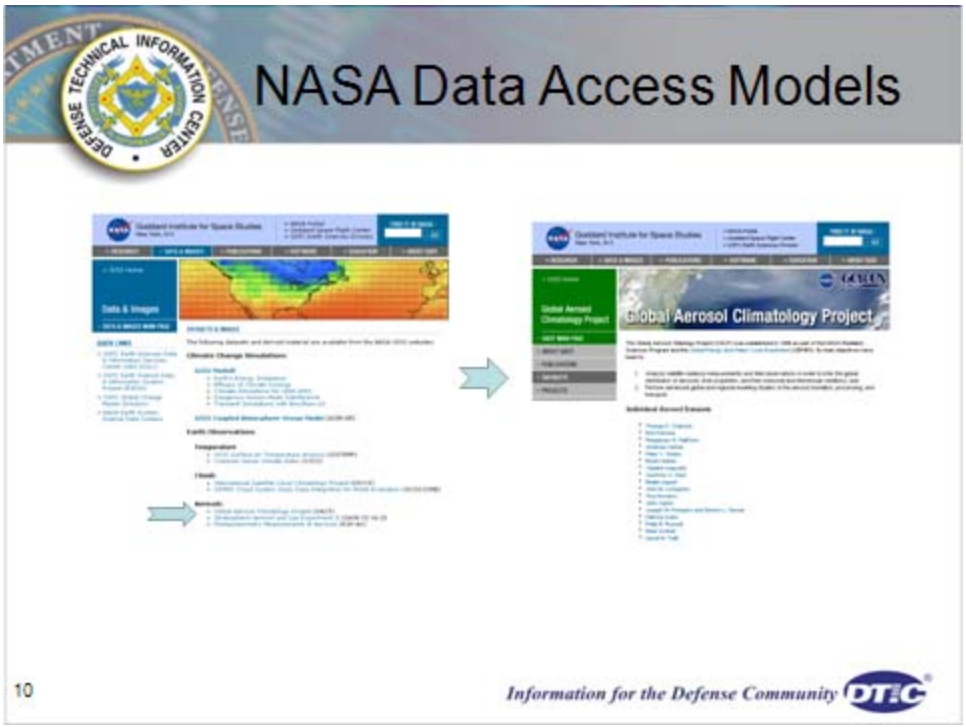
Information for the Defense Community DTIC®

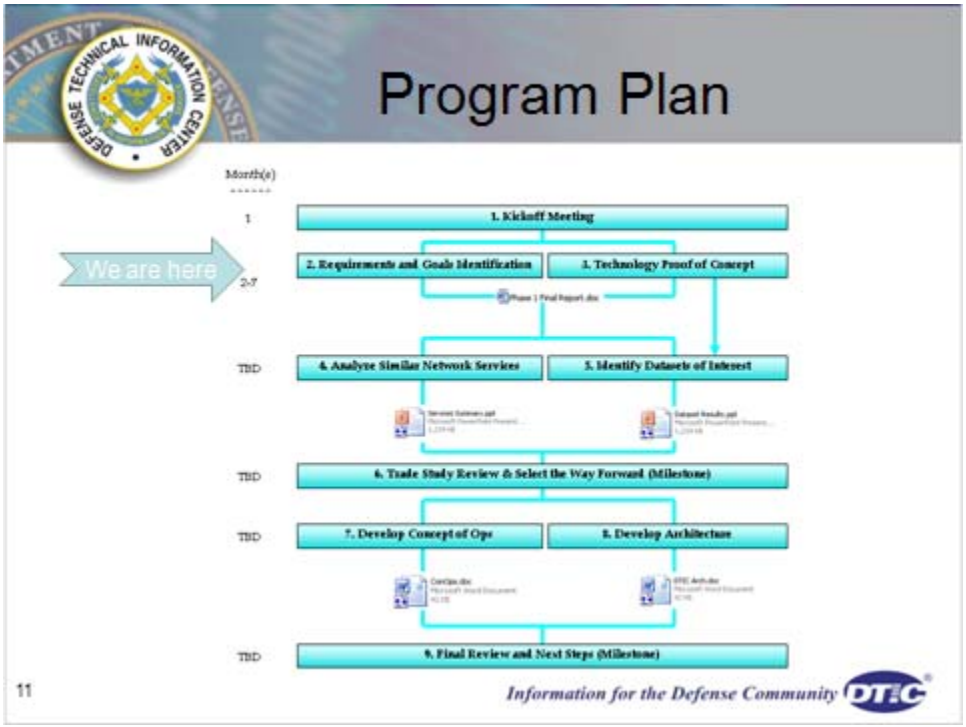
DOE & NOAA Data Access Models

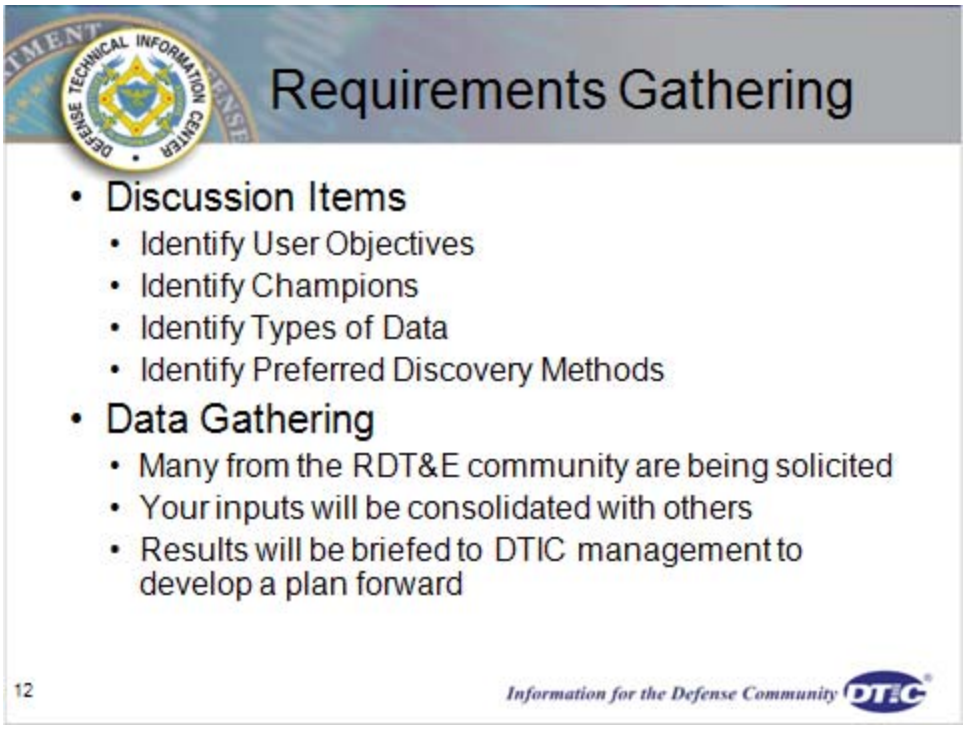


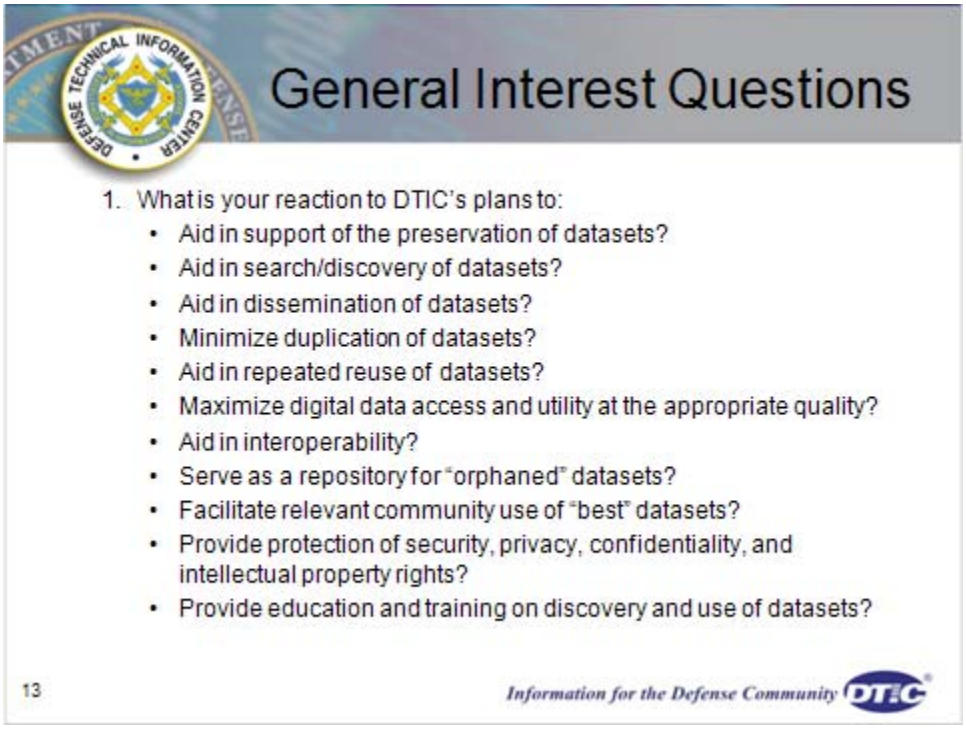
9

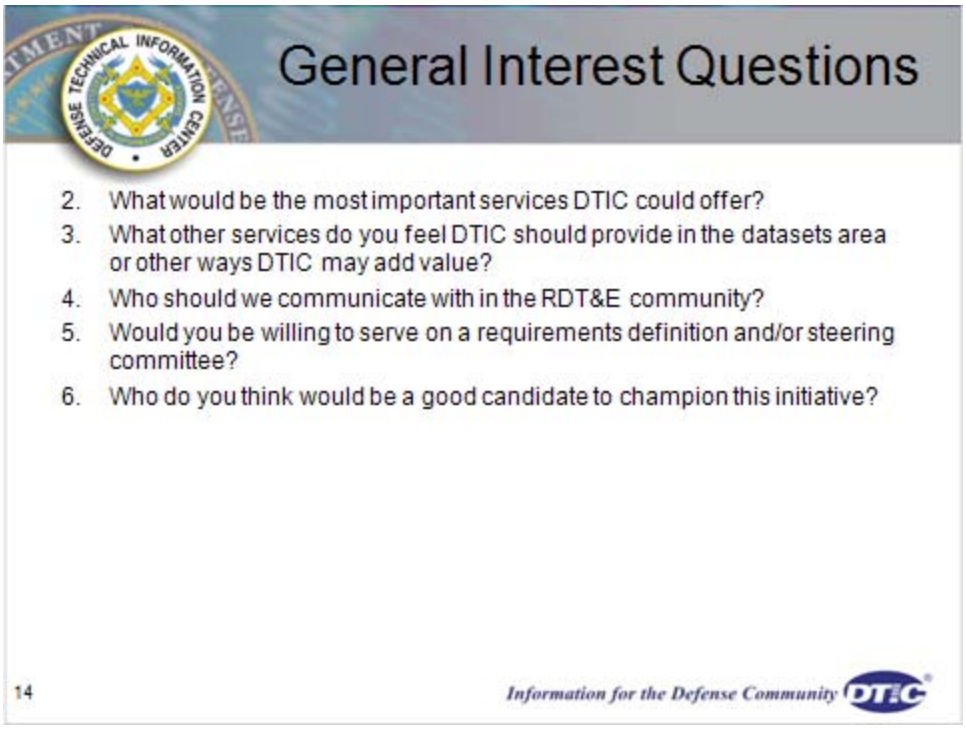
Information for the Defense Community DTIC®

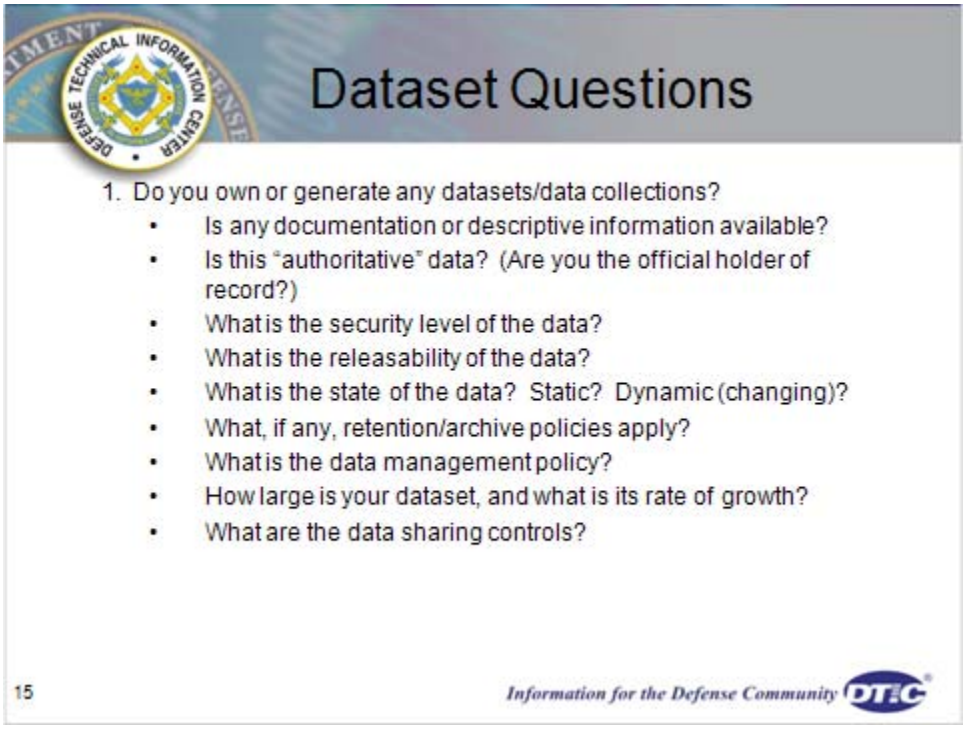

NASA Data Access Models
Global Aerosol Climatology Project
10
Information for the Defense Community DTIC

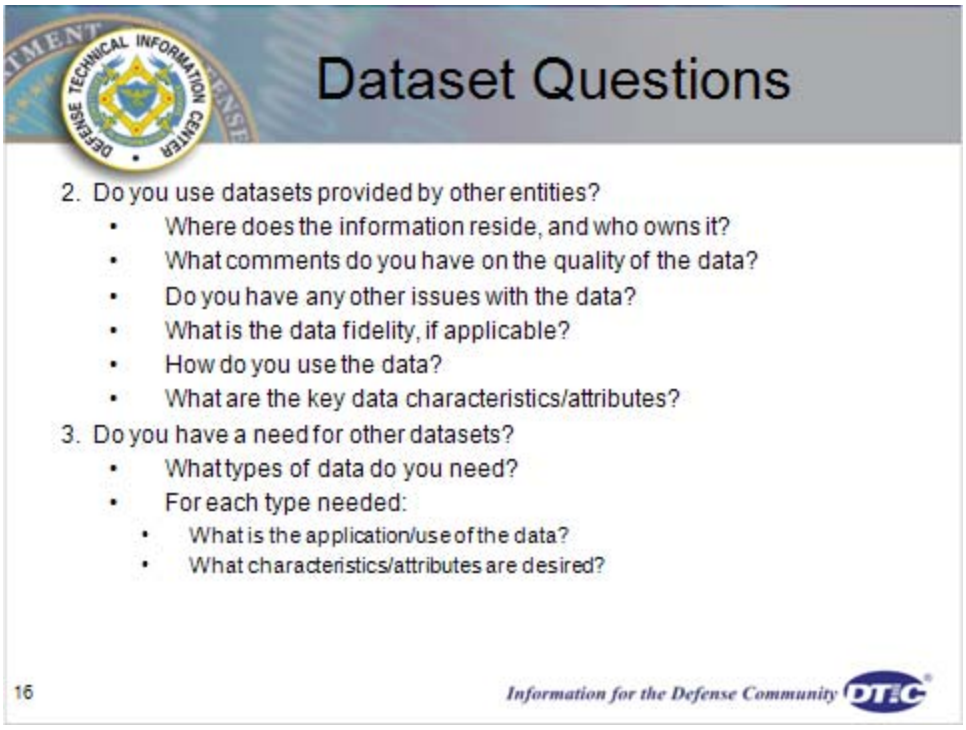

Program Plan
Month(s)
1
We are here
2-7
TBD
TBD
TBD
TBD
1. Kickoff Meeting
2. Requirements and Goals Identification
3. Technology Proof of Concept
4. Analyze Similar Network Services
5. Identify Datasets of Interest
6. Trade Study Review & Select the Way Forward (Milestone)
7. Develop Concept of Ops
8. Develop Architecture
9. Final Review and Next Steps (Milestone)
11
Information for the Defense Community DTIC

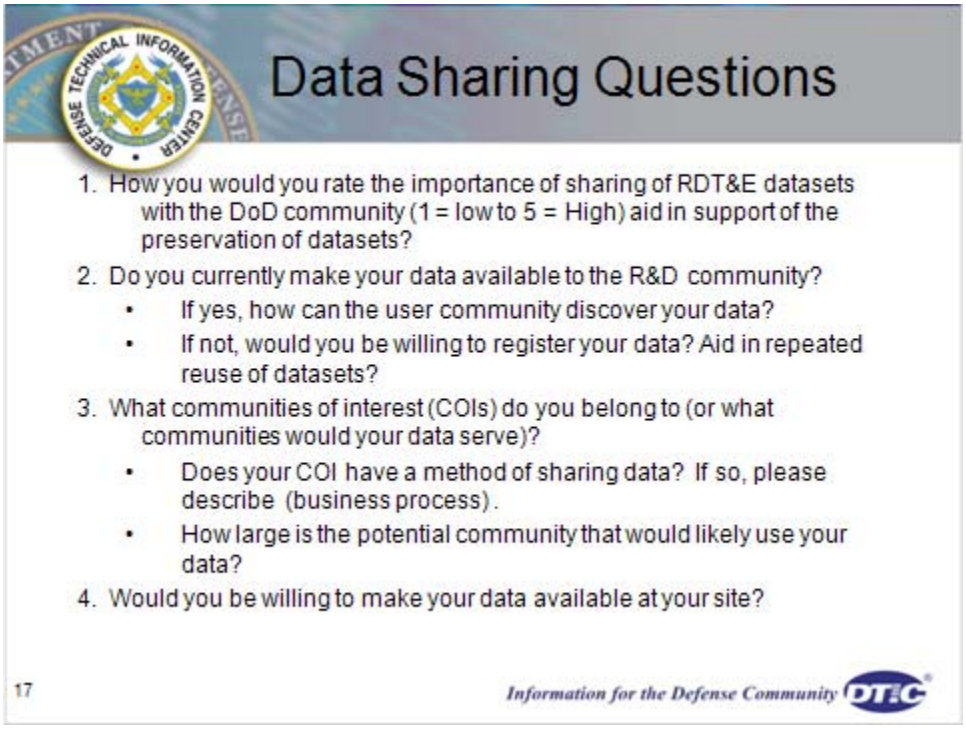

Requirements Gathering
• Discussion Items
• Identify User Objectives
• Identify Champions
• Identify Types of Data
• Identify Preferred Discovery Methods
• Data Gathering
• Many from the RDT&E community are being solicited
• Your inputs will be consolidated with others
• Results will be briefed to DTIC management to develop a plan forward
12
Information for the Defense Community DTIC


General Interest Questions
1. What is your reaction to DTIC's plans to:
• Aid in support of the preservation of datasets?
• Aid in search/discovery of datasets?
• Aid in dissemination of datasets?
• Minimize duplication of datasets?
• Aid in repeated reuse of datasets?
• Maximize digital data access and utility at the appropriate quality?
• Aid in interoperability?
• Serve as a repository for "orphaned" datasets?
• Facilitate relevant community use of "best" datasets?
• Provide protection of security, privacy, confidentiality, and intellectual property rights?
• Provide education and training on discovery and use of datasets?
13
Information for the Defense Community DTIC


General Interest Questions
2. What would be the most important services DTIC could offer?
3. What other services do you feel DTIC should provide in the datasets area or other ways DTIC may add value?
4. Who should we communicate with in the RDT&E community?
5. Would you be willing to serve on a requirements definition and/or steering committee?
6. Who do you think would be a good candidate to champion this initiative?
14
Information for the Defense Community DTIC


Dataset Questions
1. Do you own or generate any datasets/data collections?
• Is any documentation or descriptive information available?
• Is this "authoritative" data? (Are you the official holder of record?)
• What is the security level of the data?
• What is the releasability of the data?
• What is the state of the data? Static? Dynamic (changing)?
• What, if any, retention/archive policies apply?
• What is the data management policy?
• How large is your dataset, and what is its rate of growth?
• What are the data sharing controls?
15
Information for the Defense Community DTIC


Dataset Questions
2. Do you use datasets provided by other entities?
• Where does the information reside, and who owns it?
• What comments do you have on the quality of the data?
• Do you have any other issues with the data?
• What is the data fidelity, if applicable?
• How do you use the data?
• What are the key data characteristics/attributes?
3. Do you have a need for other datasets?
• What types of data do you need?
• For each type needed:
• What is the application/use of the data?
• What characteristics/attributes are desired?
16
Information for the Defense Community DTIC

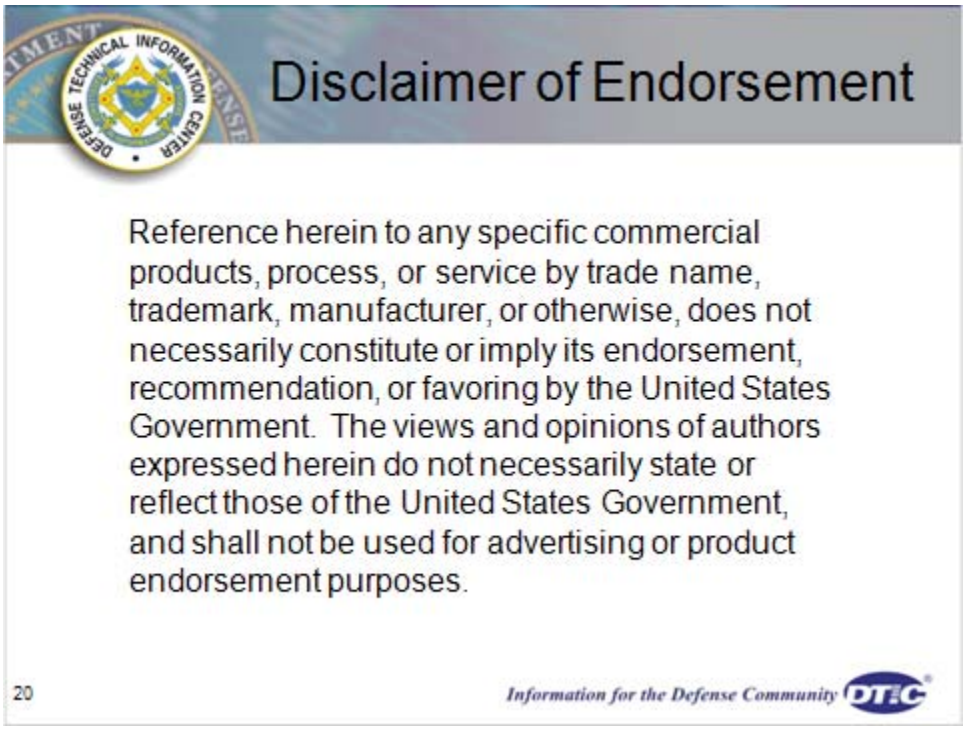

Data Sharing Questions
1. How you would you rate the importance of sharing of RDT&E datasets with the DoD community (1 = low to 5 = High) aid in support of the preservation of datasets?
2. Do you currently make your data available to the R&D community?
• If yes, how can the user community discover your data?
• If not, would you be willing to register your data? Aid in repeated reuse of datasets?
3. What communities of interest (COIs) do you belong to (or what communities would your data serve)?
• Does your COI have a method of sharing data? If so, please describe (business process).
• How large is the potential community that would likely use your data?
4. Would you be willing to make your data available at your site?
17
Information for the Defense Community DTIC

Data Sharing Questions

5. Would you prefer your data be stored and distributed by a third party?
6. Do you have sufficient metadata (scenario description, conditions, ground truth, environment parameters, measurement parameters, etc.) for the data you would share? If metadata is not available, would you be willing to create it?
7. What are your greatest challenges (barriers) in sharing data?
8. Please list the features that you would like to see in the dataset metadata repository storage structure (data attributes, accessibility, classification levels, functional capabilities, etc.) and whether the feature is a hard (must have) or soft (nice to have) need.
9. Please list the features that you would like to see in the dataset metadata repository user interface?

Contacts

- If you have any comments, questions, or would like further information please contact:
  - Mr. Gopi Nair, DTIC, gnair@dtic.mil
  - Mr. Victor Choo, ITT, vic.choo@itt.com
  - Mr. Tom McGibbon, Quanterion Solutions, tmcgibbon@quanterion.com


Disclaimer of Endorsement
Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.
20
Information for the Defense Community
DTIC

Appendix II

MSDD Stakeholders Meeting Presentation

The stakeholder organizations that participated in the MSDD meeting are listed in the following table.

Stakeholders Participating in the MSDD Meetings

Organization
Office of the Secretary of Defense (OSD)
Air Force Research Laboratory (AFRL)
Army Research Laboratory (ARL)
Naval Research Laboratory (NRL)
High Performance Computing Modernization Program (HPCMP)